TESIS DOCTORAL

APORTACIÓN A LA EXTRACCIÓN PARAMÉTRICA EN RECONOCIMIENTO DE VOZ ROBUSTO BASADA EN LA APLICACIÓN DE CONOCIMIENTO DE FONÉTICA ACÚSTICA

PRESENTADA EN LA

FACULTAD DE INFORMÁTICA

DE LA

UNIVERSIDAD POLITÉCNICA DE MADRID

PARA LA OBTENCIÓN DEL

GRADO DE DOCTOR EN INFORMÁTICA

AUTOR: AGUSTÍN ÁLVAREZ MARQUINA

DIRECTOR: PEDRO GÓMEZ VILDA

MADRID 1999

Agradecimientos

Deseo agradecer a Pedro Gómez Vilda, director de la presente tesis, los consejos, ayuda y orientación que a lo largo de estos últimos cinco años me ha venido proporcionando de manera regular.

Así mismo, quisiera reconocer la importancia que los proyectos e investigaciones desarrollados con los profesores Víctor Nieto Lluís y Rafael Martínez Olalla han tenido en este trabajo y en mi carrera investigadora.

De igual manera, no quisiera olvidar a los componentes del laboratorio de comunicación oral "Robert Wayne Newcomb" y en particular a su maestro de laboratorio José María García.

Por último, quisiera dar las gracias a todas aquellas personas que han compartido conmigo todo lo bueno y lo malo que supone realizar unos estudios tan largos y muy especialmente a mi familia, a quien dedico este libro.

With usura the line grows thick

Ezra Pound, "XLV", *The Cantos*

ABSTRACT

This thesis is based on the following hypothesis: the introduction of direct knowledge from the acoustic-phonetic field to the speech recognition problem, especially in the feature extraction step, may constitute a solid base of analysis for the determination of the behavior and capabilities of those systems and their improvement, as well.

Most of the complexity of this Ph.D. thesis comes from the different subjects related with the speech processing area. The application of acoustic-phonetic information to the speech recognition research area implies a deep knowledge of both subjects.

The research carried out in this work has been divided in two main parts: analysis of the current feature extraction methods and a study of several possible procedures about the incorporation of phonetic-acoustic knowledge to those systems.

Abundant recognition and related quality measure results are presented for 50 different parameter extraction models.

Details about the real-time implementation on a DSP platform (TMS3230C31-60) of two different parameter extraction models are presented.

Finally, a set of computer tools developed for building and testing new speech recognition systems has been produced. Besides, the application of several results from this work can be extended to other speech processing areas, such as computer assisted language learning, linguistic rehabilitation, etc.

RESUMEN

La hipótesis en la que se basa el desarrollo de esta tesis, se centra en la suposición de que la aportación de conocimiento directo, proveniente del campo de la fonética acústica, al problema del reconocimiento automático de la voz, en concreto a la etapa de extracción de características, puede constituir una base sólida con la que poder analizar el comportamiento y capacidad de discriminación de dichos sistemas, así como una forma de mejorar sus prestaciones.

Parte de la complejidad que presenta esta tesis doctoral, viene motivada por las diferentes disciplinas que están relacionadas con el área de procesamiento de la voz. La aplicación de información fonética-acústica al campo de investigación del reconocimiento del habla requiere un amplio conocimiento de ambas materias.

Las investigaciones desarrolladas en este trabajo se han dividido en dos bloques fundamentales: análisis de los métodos actuales de extracción de rasgos fonéticos y un estudio de algunas posibles formas de incorporación de conocimiento fonético-acústico a dichos sistemas.

En esta tesis se ofrecen abundantes resultados relativos a tasas de reconocimiento y medidas acerca de la calidad de este proceso, para un total de 50 modelos de extracción de parámetros.

Así mismo se incluyen los detalles de la implementación en tiempo real para una plataforma DSP, en concreto TMS320C31-60, de dos diferentes modelos de extracción de rasgos.

Además, se ha desarrollado un conjunto de las herramientas informáticas que pueden servir de base para construir y validar de forma sencilla, nuevos sistemas de reconocimiento. La aplicación de algunos de los resultados del trabajo puede extenderse también a otras áreas del tratamiento de la voz, tales como la enseñanza de una segunda lengua, logopedia, etc.

ÍNDICE

LISTA DE FIGURAS

Capítulo 1

Capítulo 2

Capítulo 3

LISTA DE TABLAS

Capítulo 5

1. INTRODUCCIÓN, ANTECEDENTES Y ESTADO DE LA CUESTIÓN

1.1. INTRODUCCIÓN

La aproximación estándar al problema del reconocimiento automático de la voz suele pasar por modificar ligeramente un sistema A, es decir, uno que responda a las premisas impuestas por el estado actual de la cuestión, para producir un sistema resultante B con la consiguiente reducción en las tasas de error entre un 1 y un 10% [BOU96].

Esta aproximación está alentada por dos factores:

1. La necesidad de retornar al menos parte de los costes en personal especializado (a veces cientos de hombres-año en tareas de investigación y desarrollo) a través de la conversión de esta tecnología actual, en productos comerciales.
2. La actividad de las agencias y organismos de investigación está condicionada, cada vez más, por la parte de explotación de los proyectos.

Aunque es lógico reducir el espacio de posibles soluciones a un problema, eliminando las líneas de investigación que aparecen como poco prometedoras, lo cierto es que si se realiza de una manera muy radical, el progreso puede constreñirse a pequeñas ganancias y las prestaciones pueden converger hacia un mínimo local del que resulte cada ver más difícil escapar [BOU96].

Por otra parte, en el campo del reconocimiento del habla siempre ha existido una tendencia por la cual, las capacidades del sistema se medían única y exclusivamente por medio de sus tasas de acierto o error. Así, desde los primeros sistemas de reconocimiento analógicos, se describían tasas de acierto del 95% [JUN96]. Esta costumbre ha fomentado dos fenómenos diferentes pero igualmente perjudiciales:

a).- La búsqueda necesaria de resultados de calidad ha hecho, que buenas ideas en forma de nuevos procedimientos o algoritmos, que incluso se han convertido en herramientas esenciales (ej. HMM), hayan sido aceptadas con tardanza. Uno de los motivos es el pequeño margen de maniobra que deja ese 5%.

b).- Incluso hoy, en sistemas comerciales se sigue reincidiendo en el mismo error.

> De esta forma, sistemas que en condiciones de laboratorio (ideales) consiguen resultados excelentes, ven como sus prestaciones decaen bruscamente cuando se los somete a situaciones o ambientes más realistas [AGA97].

En parte por todo esto, durante la última década (comparando por ejemplo con [LEE89b]), gran parte de los avances en este campo se centran más en el aumento en los tamaños de los vocabularios manejados por estos sistemas, que en un mejor conocimiento del problema. Este tipo de avance sostenido se ha visto beneficiado por la mayor disponibilidad de bases datos de entrenamiento y por la mejora de las técnicas de modelado estadístico de dichos datos [JEL97]. Como consecuencia otros aspectos del desarrollo como puede ser el reconocimiento robusto han ocupado un lugar secundario hasta hace muy poco tiempo.

De hecho, la propia definición de robusto hace referencia a todos aquellos factores que afectan negativamente a los sistemas de reconocimiento, tanto si son factores internos al problema (ej. diferencia en el habla entre distintos hablantes), como si son externos (ej. ruido acústico). Una lista con los factores que más afectan a la calidad de estos sistemas incluiría:

- Independencia del locutor. Incluye causas como edad, sexo y tímbrica de la voz.
- Ruido del entorno.
- Estados de salud y ánimo del locutor.
- Variantes dialectales.
- Uso de una lengua distinta de la materna.
- Habla afectada por condiciones de *stress* en el hablante.
- Problemas en la producción o en la percepción de la voz.

El problema del análisis del habla y de sus representaciones es complejo debido a la variabilidad presente en las señales de voz, las limitaciones de las herramientas de cálculo para con estas variaciones y sobre todo, a nuestro conocimiento muy limitado acerca del análisis y percepción del habla producida por los seres humanos [JUN96].

Sin embargo, parece evidente que el reconocimiento eficiente del mensaje hablado en modo continuo es posible puesto que nosotros realizamos esta tarea todo el tiempo y en múltiples situaciones muy diferentes entre sí. Esto parecería indicar, que el conocimiento de los procesos de comunicación entre los humanos, son la clave para resolver el problema. Así se vio durante la década de los setenta donde la mayoría de los equipos de trabajo contaban con expertos en lingüística, acústica y fonética. Pero la realidad es que esta unión no funcionó, viéndose a éstos últimos como un cuello de botella: "Cada vez que echo a un lingüista, mis tasas de reconocimiento mejoran" [JEL96].

La parametrización, es decir, la extracción de los rasgos de la señal de voz, constituye el primer paso del proceso de reconocimiento y su importancia viene dada por el simple hecho de que toda aquella información que se pierda en esta fase luego no se podrá recuperar. Es en este sentido en el que parece importante dedicar esfuerzos adicionales a esta etapa y en aplicar nuevas ideas. En este marco, la aplicación de conocimiento fonético-acústico, o al menos la aportación de información acerca de cómo se realiza este proceso en el cerebro humano, parece una propuesta razonable.

Las fases fundamentales de las que ha constado la elaboración de la tesis son las que se detallan a continuación:

- En primer lugar se ha realizado un estudio del estado de la cuestión en las áreas principales que deben ser tratadas para cubrir los objetivos propuestos para la presente tesis.
- El segundo paso fue desarrollar las aplicaciones informáticas, que implementan los diferentes tipos de modelos de extracción de parámetros empleados en el conjunto de experimentos contenidos en el trabajo. Además se desarrolló un algoritmo para el cálculo del tono fundamental de la voz, basado en procedimientos de programación dinámica.
- Seguidamente se pasó a la fase principal de la tesis, en la que ha primado la componente de investigación y donde se aplicaron los procedimientos ideados en el punto anterior a problemas concretos de reconocimiento. En total se han construido 800 sistemas de reconocimiento diferentes y a cada uno de ellos se

le ha realizado un total de 16 pruebas.

- Por último empleando parte de los desarrollos anteriores, se ha realizado una implementación en una plataforma DSP de dos versiones del sistema reconocedor para su funcionamiento en tiempo real.

La estructura por capítulos del trabajo es la siguiente:

1. En primer lugar se detallan los objetivos principales que se pretenden cubrir con la realización de esta tesis doctoral. Después en el apartado dedicado al estado de la cuestión, se recogen los trabajos realizados en el campo de la creación de sistemas de reconocimiento robusto de la voz y también los dedicados al estudio de la fonética acústica, en especial aquellos que presentan ideas potencialmente aplicables a las primeras etapas del reconocimiento del habla, como es la fase de extracción de características.

2. En este capítulo se realiza una introducción al estudio de la Fonética, tanto en su vertiente articulatoria como acústica, presentándose aquellos rasgos o características susceptibles de poder ser aplicables a los sistemas de reconocimiento de la voz.

3. El capítulo proporciona una introducción a la arquitectura de los sistemas de reconocimiento de discurso, prestando especial atención a las técnicas de tratamiento de la señal de voz que se emplean en la realización de los modelos de extracción de rasgos fonéticos.

4. En este capítulo se presenta el conjunto básico de experimentos de reconocimiento y se comentan sus resultados.

5. El capítulo introduce algunas ideas acerca de cómo poder introducir conocimiento específico a estos sistemas, así como algunas herramientas dedicadas al mismo fin.

6. Conclusiones y bibliografía.

7. Apéndices. Recogen los resultados completos de reconocimiento de los capítulos 4 y 5, así como los detalles de implementación del reconocedor de palabras aisladas en el procesador digital de señal TMS320C31.

1.2. OBJETIVOS

La hipótesis en la que se basa el desarrollo de esta tesis, se centra en la suposición de que la aportación de conocimiento directo, proveniente del campo de la fonética acústica, al problema del reconocimiento automático de la voz, en concreto a la etapa de extracción de características, puede constituir una base sólida con la que poder analizar el comportamiento y capacidad de discriminación de dichos sistemas, así como una forma de mejorar sus prestaciones.

El objetivo principal del trabajo lo constituye el estudio y análisis de la robustez de los modelos de extracción de rasgos espectrales en los sistemas actuales de reconocimiento automático del habla. En especial se analizarán seis aspectos relativos a la robustez de dichos sistemas frente a:

a).- La escasez de datos para realizar el entrenamiento o construcción de los sistemas.

b).- El dialectalismo presente en hablantes de una lengua extranjera (inglés).

c).- La distribución desigual de realizaciones de voces masculinas y femeninas.

d).- Realizaciones de habla con *stress* producido por el ruido (efecto Lombard).

e).- Realizaciones de voz que contengan ruido real con baja relación señal ruido (alrededor de 5 dB SPL).

f).- Entonaciones diferentes en las realizaciones de habla (enunciativa, interrogativa, exclamativa).

El establecimiento de un juicio crítico de las capacidades de los anteriores modelos desde el punto de vista de la fonética acústica se considera fundamental para poder progresar en este campo.

Por otra parte, el indicar cuáles pueden ser algunos posibles caminos de mejora en esta tarea puede servir para desarrollar nuevos esquemas de obtención de rasgos, que proporcionando nuevo conocimiento al problema de decodificación del habla, sean a la

vez compatibles con la configuración del resto de los módulos de los sistemas de reconocimiento. En esta línea se define un algoritmo para el cálculo del tono fundamental y que se aplica al proceso de reconocimiento.

Por último se considera importante desarrollar herramientas que permitan de manera sencilla construir y validar sistemas de reconocimiento del habla como los descritos en el punto anterior.

La completa universalidad de los resultados obtenidos en el estudio de los sistemas de reconocimiento así como su mejora en términos globales no es un objetivo de nuestro trabajo, puesto que esta meta sería lo suficientemente ambiciosa como para generar, en sí misma, nuevos trabajos de tesis. Sin embargo, se pretende ofrecer algunos principios generales que puedan ayudar a comprender el problema la complejidad del problema y a abrir nuevas vías de solución.

1.3. ESTADO DE LA CUESTIÓN

Durante las últimas décadas, la investigación en el campo del reconocimiento automático del habla se ha venido desarrollando de una forma intensa, empujada por los avances en procesamiento de señal, algoritmos, arquitecturas y plataformas de cómputo. Durante este periodo se han construido sistemas para una amplia gama de aplicaciones, que abarcan desde tareas de reconocimiento de pequeños conjuntos de palabras sobre líneas telefónicas, hasta máquinas de dictado para grandes vocabularios con capacidad para asimilar cualquier tipo de habla [RAB96].

A lo largo de este apartado vamos a revisar los principales aspectos que han condicionado y condicionan el estado actual de desarrollo de varias de las áreas que conforman el dominio del reconocimiento de voz, con especial énfasis en los aspectos relativos a la fase de extracción de características de dichos sistemas.

La historia de la investigación en el campo que nos ocupa [RAB93], [JUN96], [JUA98], se ha venido llevando a cabo durante un periodo que abarca la segunda mitad del siglo XX. Los primeros intentos por construir máquinas que realizaran tareas de reconocimiento se remontan a la década de los 50, cuando diversos investigadores trataban de explotar los principios fundamentales de la fonética acústica. En 1952, en los laboratorios *Bell*, K. Davis, R. Biddulph y S. Balashek crearon un sistema electrónico que permitía identificar para un solo hablante, pronunciaciones de los 10 dígitos realizadas de forma aislada [DAV52]. El fundamento de esta máquina se basaba en medidas de las resonancias espectrales del tracto vocal para cada dígito. Las medidas se obtenían mediante el uso de bancos de filtros analógicos.

En 1959, en la *University College* de Londres, P. Denes trataba de desarrollar un sistema para reconocer 4 vocales y 9 consonantes [RAB93]. El aspecto más novedoso de su trabajo era el uso de información estadística, acerca de las secuencias válidas de fonemas en inglés, con objeto de mejorar el rendimiento de la tarea para palabras que contuvieran dos o más fonemas. Este constituye el primer intento de incorporar conocimiento lingüístico en este tipo de sistemas.

Hasta este momento, todos los sistemas son dispositivos electrónicos. Los primeros experimentos de reconocimiento desarrollados en ordenadores tienen lugar al final de los años 50 y comienzo de los 60, principalmente en el *Lincoln Laboratory* por J. Forgie y C. Forgie [JUA98].

Durante la década de los 60, aparecen los primeros desarrollos realizados en Japón aplicando todavía, piezas de hardware específico aplicadas al reconocimiento de vocales (J. Suzuki y K. Nakata *Radio Research Lab* de Tokio, 1961), fonemas (T. Sakai y S. Doshita, *Universidad de Kioto*, 1962) y dígitos (K. Nagata, Y. Kato y S. Chiba laboratorios *NEC*, 1963) [RAB93]. Sin embargo es durante este periodo cuando se generaliza el uso de computadores en este campo.

En estos años, se inician 3 proyectos que modifican el curso de la investigación y desarrollo en el área del reconocimiento de voz de manera notable. El primero de ellos lo realizan T. Martin, A. Nelson y H. Zadell en los *RCA Laboratories* (1964) [RAB93]. Su objetivo fundamental era desarrollar soluciones realistas para los problemas asociados con la falta de uniformidad de las escalas de tiempo en los hechos de habla. Como consecuencia de este trabajo, los autores diseñaron un conjunto de métodos elementales de normalización en el tiempo, que se basaban en la detección fiable de los puntos de principio y fin de discurso. De esta forma conseguían reducir la variabilidad en las tasas de reconocimiento. T. Martin en última instancia completó el método y fundó una de las primeras compañías, *Threshold Technology*, que construía, promocionaba y vendía productos de reconocimiento de voz.

Al mismo tiempo en la Unión Soviética, T. K. Vintsyuk, propone la utilización de métodos de programación dinámica para conseguir el alineamiento temporal de pares de realizaciones de habla [JUA98]. Aunque la esencia de los conceptos relativos a la técnica de distorsión dinámica temporal o DTW (*Dynamic Time Warping*), así como la implementación de versiones rudimentarias de algoritmos para realizar reconocimiento de palabras conectadas, ya aparecían en su trabajo ("Speech discrimination by dinamic programming" *Kibernetika* 4(2), pp. 81-88 enero-febrero, 1968), lo cierto es que su difusión en el resto de Occidente no se produce hasta casi 15 años después. Para aquel entonces, otros autores ya han creado e implementado métodos más formales.

El tercer trabajo lo realiza D. R. Reddy (*Stanford University*, 1966) [JUA98] en el campo del reconocimiento de habla continua mediante el seguimiento dinámico de fonemas. La aplicación de sus ideas concluye con el reconocedor de oraciones, dependiente del hablante, para un vocabulario de 561 palabras, que aparece recogida en la tesis doctoral de P. Vicens (1969) [JUN96].

A pesar de opiniones muy críticas vertidas contra la utilidad y viabilidad de los métodos de reconocimiento automático de discurso, como los expresados en [PIE69], los años 70 representan un periodo muy activo para esta disciplina, distinguiéndose dos tipos de actividades [JUN96]:

a).- El reconocimiento de palabras aisladas comienza a ser viable y utilizable en la práctica, como consecuencia de los trabajos de V. M. Velichko y N. G. Zagoruyko en la Unión Soviética (1970), H. Sakoe y S. Chiba en Japón [SAK78] y F. Itakura en los Estados Unidos [ITA70]. Los primeros contribuyeron al avance del uso de procedimientos de encaje de patrones en el terreno del tratamiento de la voz. El grupo japonés estableció de manera formal los algoritmos, que fundamentados en la programación dinámica, podían aplicarse a la resolución de este tipo de problemas. Por último los trabajos de F. Itakura mostraban cómo los principios de las técnicas LPC (*Linear Predictive Coding*), empleadas con éxito en la codificación y compresión de la voz, podían extenderse al reconocimiento mediante el uso de medidas de distancia adecuadas sobre el conjunto de parámetros espectrales LPC.

b).- Primeros intentos de construir reconocedores de habla continua y de grandes vocabularios basados en el uso de conocimiento de alto nivel, fundamentalmente de tipo sintáctico. El fin que se persigue es poder compensar de esta forma los errores cometidos durante la fase de decodificación fonética. Muchos de estos desarrollos se realizan dentro del marco *ARPA Speech Understanding Research* (1971-1976) o inspirados por él. Los objetivos iniciales cubrían tareas de reconocimiento de oraciones para vocabularios de unas 1000 palabras realizadas por un solo hablante de manera continua. Es en este momento cuando se advierte que el conocimiento

sintáctico, semántico y contextual son fuentes de información, que permiten reducir el número de posibles alternativas que todo sistema automático de diálogo hombre-máquina debe considerar.

El sistema Hearsay I, construido por la CMU (*Carnegie Mellon University*) en 1973, como consecuencia del impulso investigador de D. R. Reddy, que a finales de los 60 pasó a está universidad, era capaz de emplear información de tipo semántico para reducir el número de posibles alternativas que el reconocedor debía evaluar [JUA98]. Así en el caso de la tarea *Voice Chess*, consistente en reconocer realizaciones de habla referidas a movimientos de una partida de ajedrez, el número de oraciones alternativas que podían producirse, se limitaba a todos los sinónimos de las posibles jugadas válidas. Sin embargo, muchas de las aportaciones de estos proyectos vienen más por la parte de la estructura software de los sistemas basados en el conocimiento o K.B.S. (*Knowledge-Based Systems*), que por los avances intrínsecos en el reconocimiento de voz [JUN96].

Otro hito durante esta década es el comienzo de los trabajos del grupo investigador de I.B.M., dedicado al dictado automático por voz para grandes vocabularios [JEL75]. Finalmente, en los *AT&T Bell Labs* (ahora *Bell Labs*, *Lucent Technologies* y *AT&T Labs-Research*), los investigadores comenzaron una serie de experimentos orientados a conseguir reconocedores realmente independientes del locutor para su uso en aplicaciones telefónicas [JUA98].

Al final de este periodo, la implementación de sistemas reconocimiento de la voz o ASR (*Automatic Speech Recognition*), se ve favorecida por la disponibilidad de tarjetas microprocesador, lo que hace posible la aparición de los primeros reconocedores a un precio bajo.

Si en la década de los 70 los sistemas de reconocimiento de vocablos aislados alcanzan una cierta madurez, los años 80 se caracterizan por la generalización en la construcción de sistemas de reconocimiento, que sean capaces de tratar con cadenas de palabras pronunciadas de una manera fluida. Encontramos durante este periodo avances muy importantes, tales como:

- La extensión de las técnicas de programación dinámica al reconocimiento de

palabras conectadas. En concreto el método en dos niveles de H. Sakoe (1979) y el método de pasada única de J. Bridle y M. Brown (1979) [JUN96].

- El desarrollo de circuitos integrados específicos que facilitan la utilización de algoritmos programación dinámica, muy costosos computacionalmente.

- El giro metodológico que se produce como consecuencia de pasar de métodos basados en comparación de plantillas a los métodos basados en modelado estadístico, debido a la extensión en el uso de los modelos ocultos de Markov o HMM (*Hidden Markov Models*) [RAB86] [RAB89]. Estos métodos habían sido desarrollados en la pasada década para tratar con problemas de habla continua [BAK75], pero su aceptación generalizada no sucedió hasta unos 10 años después. A partir de entonces se han desarrollado numerosas mejoras y actualmente constituyen los mejores modelos disponibles para capturar y modelar la variabilidad presente en el habla.

- La reintroducción de las redes neuronales [LIP87a], [LIP87b]. Los primeros modelos neuronales como por ejemplo el perceptrón, inicialmente propuesto en los años 50, volvieron a aparecer a finales de esta década gracias al desarrollo de algoritmos de aprendizaje mucho más eficaces.

- La aparición de aproximaciones al problema de la decodificación acústico-fonética para habla continua, fundamentadas en el conocimiento directo de este proceso. La tecnología de sistemas expertos había sido postulada como base para diseñar unidades de decodificación fonética que se sirvieran de la experiencia de fonetistas en tareas de interpretación de espectrogramas [CAR87].

- La grabación de bases de datos de voz como por ejemplo TIMIT (1986), que contribuyen a los avances en la disciplina y que permiten comparar los resultados entre diferentes grupos de trabajo.

Durante este mismo periodo, el programa DARPA (*Defence Advance Research Agency*), impulsó en Estados Unidos el desarrollo de mejores sistemas de reconocimiento para habla continua y vocabularios de tamaño medio y grande con independencia del locutor. Muchas de las contribuciones durante este periodo y el

principio de los años 90, provienen de los esfuerzos de la CMU a través de su sistema SPHINX [LEE89a], [LEE89b], [LEE89c].

La década de los 90 supone en cierta manera la continuidad en los objetivos ya propuestos, ampliando eso sí, el tamaño de los vocabularios a la vez que se diversifican los campos de aplicación, siendo los servicios sobre la línea telefónica uno de los que más atención acaparan en la actualidad [CHI97], [GAM97], [JUN97]. Como consecuencia de esto y del deseo de disponer de sistemas capaces de enfrentarse a situaciones cada vez más reales, ha sido en estos últimos años cuando ha surgido el interés por el estudio de los procesos de reconocimiento en condiciones de ruido y adversas en general [ACE90], [MOR94], [KAS95], [JUN96].

Algunas de las conclusiones pueden derivarse de la experiencia acumulada a lo largo de casi 5 décadas son presentadas en [JUN96]:

- ✓ Los sistemas del presente y presumiblemente los que puedan venir en el futuro se basarán al menos en parte, en modelos y técnicas que aparecieron relativamente pronto en la historia del reconocimiento automático del habla.
- ✓ La transformación de un prototipo de laboratorio de excelentes prestaciones en un sistema comercial fiable es un proceso arduo y no dominado en su totalidad.
- ✓ Las prestaciones del mejor sistema construido en la actualidad, en lo relativo a las tasas de reconocimiento, están por debajo en un orden de magnitud respecto a las que serían atribuibles al ser humano
- ✓ La solución global al problema no se encontrará de manera inmediata por el trabajo de un investigador ingenioso, sino como consecuencia de un trabajo continuado, que incorpore conocimientos multidisciplinares, incluyendo trabajos de investigación básica en los campos de producción y percepción del habla.

A la hora de afrontar la construcción de un sistema de reconocimiento podemos tomar dos enfoques principales. En el primer enfoque, los intentos pasan por encontrar lo que se conoce como propiedades invariantes o rasgos distintivos [JAK63], [FAN73]. Un conjunto de rasgos (fonéticos) distintivos, permite identificar de manera inequívoca

los componentes que caracterizan a una lengua desde el punto de vista de sus propiedades acustico-fonéticas. De esta forma, organizando este conocimiento de forma sistemática, un procedimiento de reconocimiento debería, al menos en teoría, ser capaz de identificar los sonidos presentes en una realización de habla cualquiera, por medio del etiquetado de la secuencia de vectores de rasgos. A partir de ahí, el siguiente paso sería la conversión de esa cadena de sonidos en el conjunto de palabras correspondiente.

En favor de esta aproximación está el hecho de que un experto humano, con conocimientos suficientes de los mecanismos fonéticos/fonológicos que rigen una lengua, puede deducir de manera bastante acertada los sonidos presentes en un espectrograma. Por contra está el hecho de la falta de invarianza que caracteriza al habla humana, lo que en buena medida imposibilita la búsqueda de rasgos distintivos. Además este problema se acrecienta cuando las realizaciones que estudiamos son de habla espontánea, debido fundamentalmente a fenómenos de coarticulación.

El segundo enfoque, actualmente con un claro predominio en este campo, adopta una aproximación estadística basada en encaje de patrones. De esta forma se considera que la señal de voz debe tratarse como si de un patrón estocástico se tratase.

El propósito de la fase de extracción de características es parametrizar la traza de voz en una secuencia de vectores de rasgos, que contengan toda la información relevante desde el punto de vista del reconocimiento, de los sonidos presentes en dicha realización de habla. Las propiedades que deben atesorar los citados vectores de rasgos son [RAB96]:

- Buena discriminación con objeto de poder distinguir entre sonidos similares.
- La capacidad para crear modelos estadísticos sin la necesidad de disponer de grandes cantidades de datos para el entrenamiento.
- Deben poseer propiedades estadísticas, que sean dentro de lo posible, invariantes para diversos conjuntos de hablantes y diferentes condiciones del entorno.

Hasta el momento, sin embargo, ningún conjunto de características posee todas

estas propiedades.

Los rasgos empleados en reconocimiento de voz provienen en su mayor parte de los campos del análisis y codificación de la voz. El análisis basado en la transformada de Fourier es el más empleado en la actualidad para realizar la extracción espectral en los sistemas de reconocimiento automático de la voz [RAB96]. Las implementaciones de la fase de extracción espectral incluyen (Figura 1.1):

1. Rasgos espectrales a corto plazo, calculados sobre tramas o ventanas de tamaño fijo. Aquí incluiríamos métodos de análisis como la transformada discreta de Fourier o DFT (*Discrete Fourier Transform*) y la codificación predictiva lineal o LPC (*Linear Predictive Coding*). Sobre estos esquemas es posible deducir el conjunto de coeficientes cepstrales [OPP75], [PIC93].

La técnica de predicción lineal perceptual o PLP (*Perceptual Linear Prediction*) [HER90] combina el método LPC y una aproximación a bancos de filtros por medio de la aplicación de un modelo todo-polo al conjunto de energías (o mejor dicho alturas) producidas por dicho banco de filtros. Este banco de filtros presenta un comportamiento que emula algunas de las propiedades perceptuales del oído humano. Por último calcula los coeficientes cepstrales del modelo anterior.

En los últimos años, el conjunto de rasgos espectrales a corto plazo ha sido extendido para incluir información dinámica, es decir, aquella información que relaciona el contenido espectral de una trama determinada con otras colindantes, como por ejemplo la primera y segunda derivada de n-tramas consecutivas [FUR86]. En este sentido, el cálculo de las características de segundo orden requiere de periodos más largos para una estimación fiable [APP89].

2. Propiedades espectrales normalizadas en frecuencia. En este caso se emplean escalas en frecuencia no uniformes, motivadas por un intento de emular al sistema auditivo humano que procesa la información espectral siguiendo escalas no lineales. Los ejemplo más usuales lo constituyen las escalas de mel [DAV80] y bark [HER90].

Para poder trabajar en condiciones de ruido existen algunas propuestas de parámetros como puede ser el caso de RASTA [HER92], que realiza un filtrado de banda de paso para secuencias del espectro de potencia y que proporciona un mejor resultado que el uso de técnicas de deconvolución ciega [HER93].

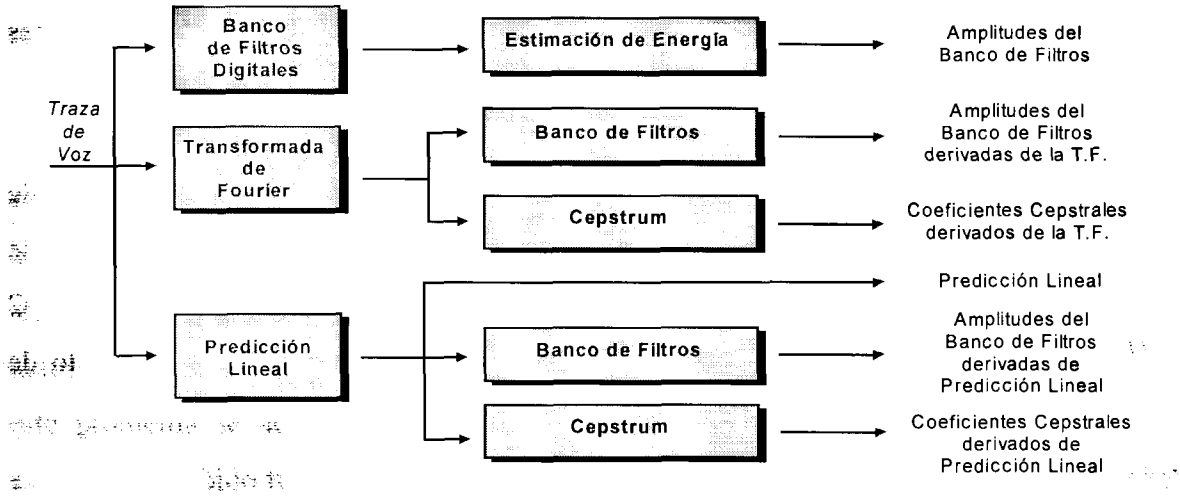

Figura 1.1. Principales grupos de algoritmos de análisis espectral básico [PIC93].

En la actualidad, encontramos diversas orientaciones a la hora de crear conjuntos de características útiles para el reconocimiento:

a).- Análisis por segmentos. El análisis a corto plazo sobre tramas de tamaño fijo, asume que la señal de voz es cuasi-estacionaria para periodos de tiempo de entre 5-30 ms. y calcula sus características con la misma resolución a través de todos los componentes en frecuencia. Este esquema es el que se sigue mayoritariamente en los sistemas actuales de reconocimiento del habla.

El análisis a largo plazo, por el contrario, realiza la extracción basándose en la información contenida en segmentos largos de habla, como puede ser una oración. Este análisis tiempo-frecuencia para periodos mayores, como por ejemplo realizan las *wavelets* es menos sensible a los artefactos causados por el uso de ventanas de tamaño fijo. Sin embargo, su utilización en el campo del reconocimiento de la voz, está todavía poco desarrollada [JUN92].

Otra manera de extraer las propiedades de un fragmento es a través de la

descomposición temporal [ATA83], que consiste en separar los elementos integrantes de la señal en términos de eventos solapados en el tiempo y la frecuencia. En este caso la clave está en integrar esta descomposición dentro del proceso de reconocimiento y de modelar el solapamiento de segmentos para habla continua.

En los últimos años destacan los trabajos realizados con parámetros multibanda. Estos métodos consisten, esencialmente, en dividir el espectro en diversas regiones de frecuencia y aplicar un procedimiento de reconocimiento de forma independiente a cada región o banda. Por último los resultados de cada parte se unifican en una solución global. Algunos ejemplos los encontramos en [MCC98], [SAR98] para habla con *stress*, o para ambientes ruidosos [OKA98]. Otros trabajos como [CER97] y [CER98] analizan el problema de la recombinación de la información de las diferentes bandas.

Encontramos también numerosos ejemplos que intentan mejorar los resultados obtenidos con vectores de coeficientes cepstrales. Así en [DOB97] se realiza el cálculo de coeficientes cepstrales con diferente resolución por banda. [PAL98] calcula un conjunto de rasgos centroides en sub-bandas, o lo que es lo mismo, medidas similares a los formantes, pero obtenidas a partir de espectros FFT y que presentan buenas propiedades frente al ruido.

Por su parte [MIL97] emplea matrices de vectores cepstrales de dimensiones entre 3 y 13 columnas de plantillas cepstrales. La matriz de cepstrum se calcula aplicando la transformada discreta del coseno a m sucesivos vectores MFCC. Con ello se pretende capturar las variaciones en la envolvente espectral a corto y largo plazo.

b).- Rasgos auditivos [LYO82]. Se basa en la consideración de que el sistema auditivo humano proporciona el método de extracción de rasgos más robusto tanto para el caso espectral como para el temporal. Algunos ejemplos relevantes incluyen medidas tales como la sincronía y la razón media [SEN88], o el histograma de intervalos EIH (*Ensemble Interval Histogram*) [GHI88]. Con el fin de incorporar estos modelos a los esquemas de reconocimiento [COH89], las propiedades calculadas se convierten al formato de rasgos a corto plazo.

Un trabajo reciente [NIT98], incorpora indicadores de funciones del sistema

auditivo dentro de un extractor de rasgos de un reconocedor de fonemas. Los resultados sobre vocales y algunos monosílabos preliminares son alentadores.

Este tipo de extracción paramétrica, que resulta ser mucho más costosa en tiempo de cómputo, proporciona mejores resultados que los sistemas basados en coeficientes cepstrales, para el caso de reconocimiento en ambientes ruidosos o cuando las condiciones de entrenamiento y prueba de los sistemas difieren bastante [HUN89]. Sin embargo, se requieren nuevas técnicas de modelado que permitan explotar de manera más adecuada los rasgos auditivos. En este sentido el conocimiento de cómo se realiza el procesamiento de la información más allá del sistema auditivo periférico puede proporcionar la solución.

c).- Rasgos articulatorios. Si el caso anterior responde a un punto de vista perceptual, ahora la motivación proviene del campo de estudio de la producción del habla. El análisis articulatorio puede ayudar al proceso de búsqueda (reconocimiento) proporcionando restricciones realistas que estén basadas en las trayectorias de estos parámetros dentro del espacio articulatorio. Este tipo de características ha visto limitada su aplicación al terreno del reconocimiento por las dificultades que conlleva asociar de manera unívoca parámetros articulatorios y señal acústica. Un intento reciente de aunar rasgos articulatorios multidimensionales en el marco de trabajo proporcionado por los modelos ocultos de Markov aparece en: [DEN94a], [DEN94b] y [ERL96].

d).- Rasgos discriminantes. Una de las metas más importantes de la extracción de características es la de encontrar aquellas propiedades que permitan reducir la tasa de errores durante el reconocimiento, con independencia de factores externos como por ejemplo el ruido ambiente. Sin un conocimiento más profundo de los fenómenos de percepción y producción del habla, no aparece claro cómo se puede determinar este conjunto de rasgos. Sin embargo, encontramos diversas aproximaciones en este terreno que se basan en el análisis de datos de habla disponibles. Una de las más importantes es el análisis lineal discriminante o LDA (*Linear Discriminant Analysis*) [HAE92], [NEY92b]. Encontramos incluso estudios de su efectividad en condiciones de ruido [SIO95].

La extracción de características discriminantes o DFE (*Discriminative Feature Extraction*) se ha propuesto como forma de obtener a partir de los datos disponibles para una determinada tarea, aquellas propiedades dependientes de dicha tarea. Este método puede además combinarse con el modelado acústico de subunidades dentro de la palabra para entrenar modelos y rasgos siguiendo una meta de mínimo error en la clasificación.

Las unidades básicas del reconocimiento son los bloques de construcción con los que se van a formar los modelos para palabras y oraciones [BAH75]. La elección de estos bloques depende esencialmente de la disponibilidad de datos de entrenamiento, es decir, la selección de un tipo de unidad implica que el material (realizaciones de voz) con el que se va a construir el modelo correspondiente es suficiente y está uniformemente repartido para todas y cada una de las unidades. Idealmente, estas unidades deben caracterizar adecuadamente los fenómenos de coarticulación presentes entre sonidos adyacentes; deben ser fáciles de entrenar y robustas a las variaciones en el habla tanto para un mismo locutor como para un conjunto de ellos, con el fin de proporcionar una tasa de reconocimiento aceptable para la tarea donde debe funcionar el sistema.

Las unidades más empleadas en los reconocedores de hoy en día son subunidades de palabra relacionadas con las distintas clases fonéticas [RAB96]. Dichas unidades se modelan acústicamente en función de transcripciones fonéticas disponibles para las realizaciones de habla. Sin embargo, por lo general no se asume como conocida la asociación exacta entre plantillas de características y subunidades. Esta asociación se establece durante la fase de entrenamiento del sistema a partir de los datos disponibles, y el resultado de todo este proceso son modelos acústicos relacionados con las entidades lingüísticas, que aparecen en las realizaciones de la entrada.

Teniendo en cuenta que la elección de dichas unidades se establece con el fin de cubrir todos los sonidos de una lengua, este patrón de reconocimiento modela virtualmente todas las secuencias de palabras posible para esa lengua. Sin embargo, en la práctica solamente se obtienen buenos resultados cuando la construcción de modelos se realiza para una tarea determinada, empleando bases de datos de voz que cubran de manera suficiente su vocabulario característico [HON92]. El entrenamiento acústico

independiente del vocabulario y de la tarea, así como el aprendizaje del vocabulario específico de esa tarea constituyen un tema de investigación importante [HON92].

La manera más simple de diseñar el conjunto de unidades fundamentales es asociar cada bloque básico de construcción con un fonema. Otras unidades incluyen pares de fonemas, sílabas, palabras e incluso frases, habiéndose probado también unidades basadas en segmentos acústicos [LEE88a].

Encontramos diferentes propuestas con objeto de incorporar información acerca del contexto lingüístico de las subunidades de palabra. Las más usuales son los tercetos de fonemas dependientes del contexto por ambos lados, conocidos en la literatura como *triphones* [SCH84], [SCH85] y más aun, los modelos que agrupan para un mismo fonema diversos contextos similares (*generalized triphones*) [LEE89b].

En la práctica, sin embargo, el número de tercetos de fonemas es excesivamente grande como para poder ser manejable. La solución usual es agrupar aquellos tercetos con contextos similares. Para ello se desarrolla un proceso automático de clasificación basado en árboles binarios o CART (*Classification and Regression Trees*) [HON91]. Con este procedimiento cada nodo no terminal del árbol representará una pregunta o característica que permitirá partir en dos el conjunto de partida y un nodo terminal estará constituido por el conjunto de datos que se empleará para construir un modelo oculto de Markov del terceto.

El criterio para partir un determinado nodo se computa como la semejanza entre los elementos de ese nodo, siguiendo una distribución usualmente Gaussiana o de Poisson [BAH91], [BAH94]. Otros métodos diferentes pero también basados en la agrupación de tercetos en clases de semejanza aparecen en: [DIG94], [GAU94], [KUB94], [ROT93].

Por último, otro ejemplo de unidad básica, aunque no muy utilizada en los sistemas de reconocimientos son los *fenones* [BAH88b]. Su característica más relevante es que constituyen un sistema auto-organizativo, ya que se extraen a partir de los datos sin presuponer ningún conocimiento fonético. Además constituyen una buena solución para incorporar nuevo vocabulario a un sistema.

Incluso en los sistemas orientados a fonemas, se suele emplear modelos específicos de palabra para aquellos casos (pocos aunque muy frecuentes) en los que la modelización de la articulación suele ser pobre, como por ejemplo para los artículos y preposiciones [LEE89b].

La tarea que se encarga de concatenar las distintas posibles palabras para formar oraciones, emplea las restricciones impuestas por lo que se conoce como gramática. Una gramática consiste en un conjunto de reglas sintácticas y semánticas que pueden especificarse como una red de estados finitos; de esta forma resulta bastante sencillo integrar directamente este conocimiento en el módulo donde se realiza la construcción de palabras. Los principales tipos de gramática son:

a).-. Gramáticas de contexto libre expresadas mediante reglas. Este enfoque es similar al que se emplea en la teoría de compiladores, aunque en este caso el proceso de selección de las reglas de producción no es determinista [ALL87].

b).- Probabilidad para cada palabra, de aparecer después de otras *N* palabras (*N-grams*), siendo *N* usualmente 2 (*bigram model*) o 3 (*trigram model*). Este enfoque se ha venido empleando de forma mayoritaria en tareas de reconocimiento de grandes vocabularios [JEL85], como por ejemplo en [AVE87], [WOO94].

El cálculo de las probabilidades se realiza mediante procedimientos de búsqueda en documentos relacionados con la tarea destino y a través de la cuenta del número de apariciones de dichos conjuntos de palabras sobre el total. Con este tipo de técnicas aparecen siempre problemas debidos a la escasez de datos y que deben afrontarse mediante técnicas de suavizado [NEY91].

Dos de las claves del éxito de los sistemas de reconocimiento del habla son por una parte los modelos ocultos de Markov, que permiten modelar las variaciones espectrales y temporales de las subunidades de reconocimiento, y el uso de técnicas de programación dinámica para búsqueda en redes, conocidas como de decodificación. Su misión es encontrar la secuencia de palabras más apropiada dentro de representaciones en forma de red finita de estados de tareas complejas [RAB89].

Encontramos numerosas variantes de los modelos ocultos de Markov:

1. Modelos de densidad discreta [LEE89].
2. Modelos continuos [LEE91]. Es el empleado por los sistemas que marcan actualmente el estado de la cuestión (ej. [YOU96]).
3. Modelos semicontinuos [HUA90].
4. Modelo de mezcla ligada [BEL90].
5. Modelos con modelado estocástico de segmentos [LEE88a].
6. Modelos con partición sucesiva de estados [TAK92].

Por último, encontramos un buen número de enfoques que emplean esquemas híbridos red neuronal/modelos de Markov [ROB94], [ZAV94], [MOR95].

Algunas de las aproximaciones al modelado acústico con modelos ocultos de Markov incluyen:

a).- Estimación de máxima semejanza o ML (*Maximum Likelihood*). La estimación de los parámetros de los HMM se suele realizar mediante esta aproximación que se basa en el algoritmo de la máxima estimación o EM (*estimation-maximization*). Otro enfoque que se ha venido usando es el ML segmentado. Aunque la estimación ML tiene buenas propiedades asintóticas, requiere de largos conjuntos de entrenamiento para alcanzar estimaciones en sus parámetros adecuadas. Por eso se han propuesto diversas técnicas de suavizado con objeto de aliviar aquellas situaciones donde la falta de datos puede ser un problema. Entre estas últimas destacan el suavizado Bayesiano y por interpolación (*deleted interpolation*) [JEL80].

b).- Estimación de máxima información mutua o MMI (*Maximum Mutual Information*). En lugar de intentar maximizar la semejanza entre los datos acústicos del entrenamiento y su transcripción en forma de modelos HMM, usando solamente los datos pertenecientes a una sola clase con el fin de entrenar al clasificador para esa clase particular, la estimación MMI aumenta la información mutua entre los datos y su correspondiente transcripción. Esta estimación toma en consideración informaciones

procedentes de datos de las otras clases debido a la necesidad de incluir las probabilidades condicionales de todas las clases en la definición de información mutua.

c).- Estimación a posteriori máxima o MAP (*Maximum A Posteriori*). Una de las formas definitivas de entrenar a las subunidades de palabra es adaptarlas a la tarea, al entorno y al locutor. El aprendizaje Bayesiano puede servir para entrenar sistemas donde, partiendo de una serie de modelos semilla (por ejemplo, dependientes del locutor) combinamos a éstos con otro conjunto de datos de adaptación, de manera que los parámetros de los modelos, así modificados, incorporen las propiedades acústicas de los nuevos datos. Esta estimación se ha aplicado con éxito tanto para incorporar nuevos hablantes en sistema dependientes del locutor como para adaptación ambiental.

d).- Estimación con mínimo error en la clasificación o MCE (*Minimum Classification Error*). Este tipo de estimación que puede emplearse también con redes neuronales, tiene como objetivo determinar el conjunto de parámetros del modelo que minimicen la tasa de error del reconocimiento basándose en datos de entrenamiento, específicos de la aplicación o de validación cruzada. Cada realización del entrenamiento se emplea para ajustar los parámetros de su modelo asociado (aprendizaje positivo) y del resto de modelos (aprendizaje negativo). En reconocedores basados en HMM, se han venido aplicando algoritmos de descenso probabilístico generalizado o GPD (*Generalized Probabilistic Descent*), en la estimación de parámetros de modelos siguiendo una orientación basada en criterios de mínimo error [CHO92]. Este tipo de enfoque MCE/GPD es capaz de maximizar la separación entre modelos de unidades de habla de forma que, la discriminación y robustez del sistema reconocedor asociado a ellos puede mejorarse de manera simultánea.

La señal de voz es en esencia una señal dinámica que puede presentar muchas fuentes de variación. La producción de diferentes sonidos conlleva movimientos articulatorios muy diferentes, a la vez que existe bastante libertad en la temporización y grado de apertura de los movimientos del tracto vocal. El hablante puede variar su producción del habla hasta un cierto grado y seguir transmitiendo el mismo mensaje. La (Tabla 1.1) muestra una posible clasificación de las categorías en que podemos clasificar a los diferentes estilos de habla [JUN96].

Aunque no siempre es sencillo delimitar la diferencia entre diferentes estilos (ej. habla normal y cuidadosa [OHA95]), en general se puede comprobar como la mayoría de estas variaciones, si bien son beneficiosas a la hora de aumentar la naturalidad de los sintetizadores de voz, perjudican el rendimiento de los sistemas de reconocimiento [JUN93].

Variaciones en el estilo	Velocidad de producción	*Stress*	Contexto	Cualidad de la voz
Cuidadoso	Normal	En condiciones de ruido (e. Lombard)	Deportivo	Aspirada
Claro	Despacio	Factores emocionales	Profesional	Ronca
Articulado	Rápido	Condiciones de carga cognitiva	Entrevista	Susurrante
Formal	Muy rápido		Conversación libre	Tensa
Casual			Diálogo hombre-máquina	Laxa
Espontaneo				Modal
Normal				
Leído				
Dictado				

Tabla 1.1. Clasificación general de las variaciones del habla. Para cada categoría (columna) se muestran sus posibles valores.

Un aspecto importante es la correlación existente entre estilos de habla y su correlato acústico. La producción de habla clara, por ejemplo, no es equivalente a producir habla normal con más volumen, sino que en realidad conlleva la reorganización de patrones acústicos y gestos articulatorios [JUN93]. De igual forma, el habla producida en condiciones de ruido o el habla articulada claramente aumenta su duración en el tiempo [JUN95]. En términos de inteligibilidad también encontramos diferencias. Así el habla clara [REM91] y la producción de conjuntos de palabras poco propensas a ser confundidas pero en condiciones de ruido [SUM88], resultan ser más inteligibles que su equivalente producido con un estilo normal y en condiciones silenciosas. Por el contrario, si las palabras son fácilmente confundibles la inteligibilidad para la situación con ruido es menor [JUN93].

El tipo de datos que se emplean en las investigaciones de los sistemas automáticos de reconocimiento del habla pueden clasificarse en dos grandes grupos: habla producida a partir de la lectura de textos previamente preparados y habla espontanea (mucho más difícil de tratar). Algunas diferencias notables entre ambos estilos son el predominio de las disminuciones del tono fundamental junto a una mayor duración y número de pausas en el primer caso frente a una mayoría de elevaciones para el segundo. Otra gran diferencia entre ambos estilos proviene del fenómeno de la vacilación. En la producción del habla espontanea, el locutor a la vez que pronuncia debe elaborar el mensaje, es decir, la carga cognitiva es mayor para este último caso. Una de las consecuencias de este fenómeno es que junto al mensaje lingüístico encontramos pausas y partes extrañas. La duración y el tono fundamental son buenas claves incluso en entornos ruidosos para la distinción entre pausas, producidas en los bordes sintácticos y aquellas otras internas [OSH92].

El estudio y tratamiento de las emociones y el *stress* en los procesos de reconocimiento de discurso resulta difícil, entre otras razones, por la dificultad para poder realizar experimentos bien controlados [BAB95]. Las fuentes de alteración del habla por este capítulo son numerosas y diversas siendo las principales: entorno, emociones y carga de trabajo. Dentro del entorno podemos encontrar causas debidas al ruido, a factores mecánicos como aceleración y vibración y a agentes físicos como el calor. De todos los factores que inducen *stress* en el hablante el más estudiado se corresponde con el llamado efecto Lombard [JUN93], puesto que resulta muy sencillo producir el estímulo (ruido) y separar sus efectos de las causas que lo producen.

Una de las razones de la escasez de conocimiento es este campo tiene su origen en las grandes diferencias en las respuestas, que ante el mismo estímulo producen personas distintas. El indicador más sensible de la presencia de emociones en realizaciones de habla viene dado por el seguimiento del contorno del tono fundamental [HEC68]. Estudios acerca de emociones simuladas tales como, rabia, miedo o pena, muestran las diferencias presentes en el contorno de la frecuencia fundamental, espectro promediado, características temporales y precisión en la ejecución articulatoria [WIL72].

Dentro de las diferencias entre locutores, el aspecto más relevante es la variación

en el habla de hombres y mujeres. De media, los hablantes femeninos presentan una estructura de formantes más alta en frecuencia y más presencia de fenómenos de aspiración Sin embargo, existen muchas diferencias entre individuos del mismo género [KLA90]. En general las mujeres tienden a articular más que los varones [HEN92]. Esto se debe en parte a la menor velocidad de producción del habla, que constituye un importante factor para aumentar la inteligibilidad [KLA90]. Otro factor que se aprecia es que el triángulo vocálico de las voces femeninas se extiende más allá que el correspondiente a las voces masculinas, en particular para el F_1 [HEN92]. Además existen algunas diferencias básicas en los mecanismos de fonación y articulación para los dos grupos [TIT89]. La diferencia entre la laringe de hombres y mujeres es la responsable de las diferencias en el valor medio de F_0, flujo de aire medio y en la potencia aerodinámica producida. También los estudios en condiciones de ruido sugieren diferente respuesta entre ambos grupos [JUN93]. La variabilidad para un mismo hablante, debida al aumento de esfuerzo vocal cuando el sonido se produce con ruido de fondo, tiende a ser mayor en hablantes femeninos.

Controlar las individualidades del habla o añadir determinadas claves a los algoritmos ha sido el enfoque que se ha llevado a cabo en campos como la síntesis de voz, como forma de mejorar la naturalidad cuando se desea convertir la calidad de la voz de un hablante en otro [ABE88], [ABE91].

Uno de los aspectos, que ha atraído más interés a través de los años entre los investigadores, es la determinación de los límites de la variabilidad del habla y de la existencia de propiedades invariantes. Es evidente que existen muchas fuentes de variación en el proceso de producción de la voz. Unas tienen que ver con el locutor y otras con el canal de comunicación. Todos estos factores interaccionan con la inherente redundancia de la señal de voz, su contenido lingüístico y el contexto a la hora de determinar la inteligibilidad del mensaje hablado. Muchos de los orígenes de la variabilidad son difícilmente controlables incluso en entornos de laboratorio, donde algunos elementos tales como micrófonos, instrumentación y/o la tarea pueden producir *stress* en el sujeto.

Los aspectos relativos a la invarianza han sido largamente estudiados. Algunos

investigadores han propuesto teorías acerca de la invarianza en el proceso de producción del habla en el sentido de que la señal de voz está fuertemente estructurada y contiene patrones acústicos invariables que pueden formar parte de vectores de rasgos [BLU79]. Estas propiedades permanecerían invariantes a lo largo de diferentes locutores, contextos fonéticos y lenguas. El concepto de invarianza en sentido estricto, sin embargo, es en cierta manera sorprendente, considerando el hecho de que no existe un acuerdo, incluso entre los lingüistas, en la clasificación y definición exacta del material fonético.

Este hecho, no obstante, no implica el que no podamos encontrar características o relaciones entre distintos rasgos, que puedan servir de claves a la hora de determinar la estructura acústica de una realización de habla, al mismo tiempo que reducen la variabilidad ya sea de tipo intra-locutor o inter-locutor. Así [POR81] defiende cómo la relación entre duración del cierre y la duración de la vocal proporciona una pista relativa a la propiedad de sonoridad que es contextualmente invariante. La proporción se mantiene a pesar de la velocidad de producción del habla, número de sílabas presentes en las palabras de prueba y la tensión de las vocales. Estos rasgos relacionales también han sido propuestos para situaciones en presencia de habla afectada por ruido [JUN93]. Las derivadas temporales y los coeficientes cepstrales aparecen también como formas de mejorar el reconocimiento para este último caso [HAN90].

Dentro de las fuentes de variación introducidas por los propios locutores, podemos establecer una clasificación en dos grandes grupos:

a).- Variación intra-locutor. Dos de las fuentes principales de variación personal están directamente relacionadas con el estado psicológico y fisiológico del hablante y con el objetivo que se persigue durante el proceso de comunicación. Estas variaciones son las que se conocen como estilos de habla y varían según el momento, de acuerdo a diversos factores como son el entorno y el contexto.

El resultado de estos factores suelen ser pequeños cambios en el nivel articulatorio que sin embargo, suelen reflejarse como grandes variaciones en el nivel acústico.

Para el caso de las variaciones producto de cambios en el ritmo de producción de

la voz, las frecuencias de formantes vocálicos se ven más alterados que para el caso de consonantes. Además la variación no es uniforme, aunque por lo general el formante que más varía es el F_2 y el que menos lo hace es el F_3. En realizaciones lentas la duración de las vocales aumenta mientras que en habla rápida la relación de duración vocal y consonante se mantiene [KUW97].

Las variaciones debidas a fenómenos del entorno pueden clasificarse en dos grupos [JUN96]:

1. Estáticas, como la acústica de las habitaciones, reverberaciones, características de los equipos de registro, elementos de captura de señal (micrófonos/teléfonos), la personalidad del locutor y sus características físicas.
2. Dinámicas, como el ruido de fondo, el *stress*, la carga cognitiva y la posición de los micrófonos.

Los cambios en la configuración del entorno son difíciles de controlar y reproducir para muchas de las citadas fuentes de variabilidad. Encontramos numerosos estudios que tratan con algunos de estos aspectos en tareas de reconocimiento de voz:

- Estudio en ambientes con fuerte reverberación, empleando conjuntos de micrófonos y técnicas de conformación de haces (*beamforming*) [VAN90].
- Normalización espectral para compensación de ruido y habla con *stress* [HAN95], [HAN96]
- Investigaciones acerca de la degradación en las tasas de reconocimiento como consecuencia de las diferencias presentes en las condiciones de entrenamiento y prueba de los sistemas de reconocimiento [ACE90].
- Estudios acerca de la dependencia/independencia de los micrófonos en la calidad del reconocimiento [ACE90], [ANA94], [DAS94], [LIU94].

Por su parte, las variaciones debidas al contexto pueden agruparse según éste en:

- Contexto fonético local o coarticulación.
- Contexto lingüístico, incluyendo conocimiento sintáctico, semántico y

pragmático de la situación.

- Interacción social.

Las dos primeras clases de variación son el fruto de las oraciones producidas en un determinado momento así como de su relación dentro del discurso hablado y la tercera está relacionada más con el ambiente, aunque no físico. Encontramos un conjunto de estudios relativos a los cambios en el habla como consecuencia de la interacción social y datos acerca de los cambios en la realización de vocales [LAB86]. Los resultados muestran que existe una variación notoria en los valores medios de frecuencias de los formantes para los diferentes casos.

b).- Variaciones inter-locutor. Las diferencias fisiológicas entre personas son una de las causas más importantes que determinan las diferencias entre el habla de dos individuos diferentes. Las principales diferencias fisiológicas las encontramos en la longitud y forma de las cuerdas vocales, la geometría del tracto vocal y la configuración del tracto nasal. Las diferencias en el tracto vocal para hombres, mujeres y niños produce que las frecuencias de los formantes no estén relacionadas linealmente. Un tracto vocal corto producirá, por lo general, valores más altos para los formantes. Más aun, el aparato fonador de la persona evoluciona con la edad.

[BUS95] realiza un estudio que intenta analizar las variaciones en el habla como consecuencia de la edad y el género de niños de entre 5 y 11 años. Las conclusiones son que la altura de los tres primeros formantes se hacen menores a medida que aumenta la edad, siendo las posiciones frecuenciales de las niñas mayores. En cuanto al tono fundamental se observa que varía de forma inversamente proporcional al incremento en la edad, no apreciando por lo demás diferencias significativas entre los dos grupos.

La diferencia en el habla de las personas como consecuencia de su edad es un factor importante que limita las prestaciones de los sistemas de reconocimiento [WIL96].

Por otra parte, los diferentes hábitos articulatorios también contribuyen a aumentar la variabilidad entre locutores, siendo función de la personalidad del hablante (estilos de

habla usuales), por una parte y por otra, fruto de las peculiaridades dialectales del grupo al que pertenece la persona.

A pesar de la anterior clasificación, nos encontramos con que algunos factores también alteran a otros. Las condiciones del entorno, por ejemplo, introducen variaciones en los contextos fonético y lingüístico. El entorno, además produce variaciones que pueden ser diferentes o aparecer en distinto grado para diferentes personas. Incluso en pruebas de laboratorio, donde se intenta controlar de la mejor manera posible todas las variables anteriores, es difícil anular la variación introducida por la predisposición y el nerviosismo de los locutores, el formalismo que el propio experimento induce en ellos en forma de corrección de estilo, etc.

En general, durante el pasado se ha venido considerando a la variabilidad como una fuente de ruido [PIS97]. Sin embargo, la propia variabilidad puede proporcionar información útil acerca de las características del hablante o de su contexto fonético. De hecho algunos estudios se centran en los mecanismos de normalización que los locutores humanos producen al hablar [HOL86].

Para tratar con este problema [HUA93], encontramos un buen número de propuestas de normalización de vocales en función de la longitud del tracto vocal [WAK77], [MIL89], [WEL98b].

Durante la última década y dentro del campo del reconocimiento de la voz, se han realizado progresos sustanciales en la adaptación dinámica de los sistemas de reconocimiento a nuevos hablantes, a través de la modificación o normalización de las representaciones fonéticas con objeto de reflejar las características acústicas de los hablantes de forma individualizada [SCH87]. Un enfoque usual es normalizar el espectro de frecuencias, a partir de las informaciones extraídas para los valores de alguno o algunos de los tres primeros formantes [CLA97], [GOU97], [LIN97], [LEE98].

La adaptación a un nuevo hablante mediante la modificación o adaptación de los parámetros de los modelos ocultos de Markov [ZHA97] es otra posibilidad estudiada, especialmente interesante para sistemas de dictado o dependientes del locutor.

La normalización de la velocidad de producción del habla y su relación con técnicas de reconocimiento de voz ha dado lugar a numerosos estudios como por ejemplo: [MAR98], [PFA98].

Los efectos del ruido en las personas son numerosos. Uno de los más obvios es el que afecta al proceso de comunicación, al enmascarar parte de los contenidos del mensaje hablado. En los sistemas de reconocimiento automático de discurso, las prestaciones se mantienen uniformes para relaciones señal ruido de 25dB o superiores, pero existe una fuerte degradación de éstas conforme el nivel de ruido aumenta. El problema radica en que los métodos de análisis de la voz son muy sensibles a las variaciones que el ruido introduce en la señal.

El ruido inducido por el entorno o por los canales de comunicación puede ser lineal en el dominio del espectro de potencia (aditivo), lineal en el espectro logarítmico o cepstral (convolutivo) o simplemente, no lineal en ambos dominios, siendo este último caso es el más difícil de tratar. Los ruidos causados por el entorno son usualmente de tipo aditivo, pero no siempre es el caso. Sin embargo, la mayor parte de las investigaciones sobre reconocimiento de voz en entornos ruidosos se han realizado empleando ruido blanco Gaussiano o también ruido rosa sumado a la señal [JUN96]. Otra restricción generalmente empleada es la que hace referencia a la estacionaridad y la incorrelación del ruido respecto a la señal de voz. No obstante, la consideración de la estacionaridad a largo plazo excluye distorsiones frecuentes, por ejemplo, en los entornos de oficina (estornudos, portazos, etc.).

Los tipos de ruido típicos se pueden clasificar en diversas categorías de acuerdo a la tarea y el contexto en que tiene lugar la conversación. Una clasificación genérica permitiría agrupar los ruidos que afectan a los sistemas de reconocimiento del habla en [JUN96]:

a).- Entornos de oficina.
b).- Militares e industriales.
c).- Transporte.
d).- Canal. d.1).- Debidas a línea telefónicas.
d.2).- Distorsiones inducidas por el micrófono.

d.3).- Efectos acústicos debidos a ecos y reverberaciones.

Los ruidos también pueden distinguirse de acuerdo a sus distribuciones tiempo-frecuencia:

a).- Periódicos.	Ej. ruido de motores
b).- Impulsivos.	Ej. disparos o descargas eléctricas.
c).- De banda estrecha.	Ej. ruido de un coche.
d).- De banda ancha.	Ej. ruido de otras conversaciones cercanas (*cocktail party effect*).

Para los procedimientos de reconocimiento, el ruido de banda ancha es el más difícil de manejar.

El ruido de oficinas es aditivo en esencia y se caracteriza por concentraciones de energía sobre ciertas partes del espectro. Existe una gran cantidad de fuentes de ruido, tales como, ventiladores, fluorescentes, máquinas de escribir, ordenadores, conversaciones, movimientos de sillas, portazos, etc. Los ruidos de tipo periódico que provienen de ventiladores o de interferencias eléctricas pueden ser suprimidos por medio de filtrado, al poder ser estimados adecuadamente.

Los ruidos de tipo industrial y militar suelen ser aditivos y se caracterizan por ser de banda ancha pero con fuertes componentes a determinadas frecuencias. En general, estos ambientes son los que presentan mayores niveles de ruido como por ejemplo los valores medidos en la cabina de un F16 (85dB- 112dB) [RAJ86]. En este tipo de entornos el fuerte *stress* inducido es también una causa de la falta de robustez que presentan los reconocedores de voz.

Los ruidos asociados a medios de transporte (automóviles, trenes y aviones) dependen de diversos factores propios de cada uno de ellos. Para el caso de automóviles [OH95] encontramos: motor, ventiladores, transmisión, ruedas y efectos aerodinámicos. En trenes el ruido es fruto del movimiento, vibraciones mecánicas y los raíles. En aviones el ruido aerodinámico y de combustión, vibraciones, ventiladores y compresor son los orígenes principales del ruido. En todos los casos los pasajeros pueden añadir su contribución como por ejemplo en coches el equipo de música.

Aparte de aplicaciones militares, el ruido en coches ha atraído el mayor interés debido fundamentalmente al potencial mercado disponible para aplicaciones de reconocimiento de voz [OH95], [FIS98], [LEC89].

En automóviles, podemos diferenciar entre ruido externo (el producido por el coche propiamente dicho y su movimiento), e interno (producido por el pasajero). La mayoría de los estudios se ha concentrado en el ruido externo porque resulta más sencillo de reproducir. Este ruido es aditivo en el dominio del espectro de potencia y consiste en bajas frecuencias, esencialmente debidas a fuentes mecánicas (motor y neumáticos) y a ruidos de espectrograma plano producidos por fenómenos aerodinámicos. El ruido puede llegar a ser intenso cayendo la relación señal ruido a -5dB para un micrófono situado en el salpicadero cuando el coche está en movimiento con las ventanillas subidas [LEC89].

Entre las distorsiones de canal, las más estudiadas son las relativas a la línea telefónica, disponiéndose de numerosas bases de datos específicas. Las características de este medio que más dificultan la calidad de la voz son: el ancho de banda limitado, el tipo y la calidad del teléfono y por último, las variaciones en el ambiente de funcionamiento. Por su parte, el ruido estacionario aditivo, el ruido impulsivo, los tonos de baja frecuencia y las diferencias de respuesta entre bandas de frecuencias son los factores que más afectan al reconocimiento en el caso de ruido telefónico [MOR94].

Otro tipo de distorsiones de canal son aquellas introducidas por los micrófonos y por la acústica de las habitaciones en forma de ecos y reverberaciones. Estos dos casos pueden verse también como fuentes de variabilidad.

Los sistemas de reconocimiento de voz funcionan peor cuando los micrófonos empleados durante la construcción del sistema son diferentes a los que se emplean durante la fase de prueba de éste [ACE90]. De hecho un micrófono actúa como un filtro lineal de la señal de voz y presenta diferentes grados de envolvente espectral dependiendo de sus características; por tanto, este tipo de distorsión se convoluciona con la señal. Cada tipo de micrófono es más o menos sensible al ruido de fondo y además la distancia entre micrófono y locutor introduce diferentes niveles de ganancia.

Las habitaciones con paredes sólidas aumentan la reverberación de tal manera que el sonido se prolonga y decae con el tiempo en una proporción logarítmica que en la práctica suele ser irregular. El caso de los teléfonos manos libres, cuando el locutor está lejos del micrófono es otro ejemplo de este fenómeno. Este tipo de ruido es de tipo convolutivo.

Los efectos del ruido se reflejan en el espectro por una disminución en las diferencias entre picos y valles, aunque son estos últimos los que más se modifican. El ruido tiende a reducir el rango dinámico y la varianza entre tramas. En el nivel del reconocimiento se observa como las características dinámicas se ven menos afectadas que las instantáneas en ambientes ruidosos [JUN96].

Los humanos para reconocer el habla en condiciones de ruido, emplean diversos recursos como son: la audición binaural, la información visual así como su conocimiento del contexto y la lengua. La adaptación al ruido es otro fenómeno que contribuye a la capacidad de los humanos para entender voz ruidosa. En este sentido encontramos diferentes estudios que tratan de determinar los aspectos a los que personas con deficiencias auditivas son más sensibles [BOO96], [LEE96].

Actualmente muchos sistemas de reconocimiento asumen para su funcionamiento condiciones silenciosas o como mucho modelan el ruido como de tipo Gaussiano o rosa, que no representa unas condiciones realistas. Consecuentemente se producen reducciones dramáticas en las tasas de reconocimiento entre sistemas probados en el laboratorio y en condiciones reales [AGA97]. Con objeto de resolver el problema del reconocimiento en entornos ruidosos se han investigado tres aproximaciones [EPH87]:

1. Mejorar los sistemas que funcionan adecuadamente en condiciones de laboratorio.
2. Diseñar nuevos sistemas más robustos a las condiciones ruidosas.
3. Entrenar al sistema en un ambiente similar al que luego se va a utilizar.

En la práctica lo anterior se resume en dos aproximaciones principales [COL97]:

a).- Adaptación dinámica de parámetros y estimación óptima de parámetros.

Muchas técnicas robustas al ruido se fundamentan en modelos estadísticos, que caracterizan las diferencias entre el habla usada en la fase de construcción del sistema y la que realmente aparece durante el funcionamiento de éste. Los modelos típicos de adaptación a la variabilidad acústica asumen que la voz está corrompida, bien por ruido aditivo con un espectro de potencia desconocido [POR84], [GAL92], [LOC92], [YAN95], o por una combinación de ruido aditivo y filtrado lineal [ACE90].

Una aproximación interesante, cuando el ruido es conocido, consiste en utilizar información extraída del ruido de fondo a partir de ejemplos con el objetivo de transformar algunas propiedades estadísticas, por ejemplo medias y varianzas de los modelos fonéticos que han sido desarrollados para habla limpia y permitir a estos modelos caracterizar la voz con el ruido [VAR90].

Algunos algoritmos robustos se basan en aplicar vectores de corrección sumados a las características extraídas de la señal de voz. Las tasas de reconocimiento pueden mejorarse sustancialmente permitiendo que los vectores de corrección dependan de la relación señal/ruido, localización específica dentro del espacio de parámetros para ésta última o la supuesta identidad de un fonema [NEU94].

b).- Empleo de varios micrófonos. Pueden obtenerse mayores mejoras en las tasas de reconocimiento cuando se emplean varios micrófonos en condiciones de fuerte ruido. Estudios llevados a cabo en entornos de oficina [SUL93] confirman que el uso de métodos de conformación de haces (*beamforming*) del tipo retardo y suma en combinación con un algoritmo de postproceso, que compense la coloración espectral introducida por el conjunto de micrófonos, puede reducir los errores en más de un 60%.

Una conjunto de micrófonos que haga uso de técnicas de filtrado adaptativas minimizando el error cuadrático medio, puede funcionar bien cuando la degradación de la señal está dominada por ruido aditivo, pero no funcionan bien en entornos reverberantes, donde la distorsión es al menos en parte la versión retardada de la señal de voz deseada [ALV90]. Una posible solución a este último caso es la de adaptar el sistema solamente en los periodos de ausencia de voz [VAN90].

Una tercera aproximación cuando se usan varios micrófonos, es la de emplear

algoritmos basados en la correlación cruzada, que tienen la habilidad de reforzar los componentes del campo sonoro, que provienen de un cierto ángulo. Estas técnicas, si bien son muy atractivas porque emulan el comportamiento binaural del oído humano, han demostrado solamente una superioridad muy pequeña respecto a las aproximaciones con conformación de haz [SUL93].

Diversos estudios se han ocupado de los aspectos relativos a la percepción de las distinciones fonéticas en presencia de ruido. En general, muestran cómo existen distinciones más robustas que otras. En el caso de las consonantes, el carácter nasal y el oral, que se basan en pistas proporcionadas por las bajas frecuencias, son los más robustos [MIL55], [WAN73]. Para las vocales [PIC57], las distinciones basadas en el primer formante son más robustas que aquellas otras que se fundamentan en el segundo formante o la duración. Sin embargo, estos resultados dependen de las características espectrales del ruido.

No todos los parámetros son igual de fiables cuando se identifican en condiciones ruidosas. El contorno del tono fundamental, la trama de máxima potencia, las regiones vocálicas y la separación temporal entre sílabas tónicas se encuentran entre las características más resistentes al ruido. La utilización de este conocimiento permite diversas aplicaciones como son la detección robusta de unidades silábicas [MER75] o la detección de principio y fin [LAM81] fiable en condiciones de ruido [JUN94].

El reconocimiento y compresión del habla son básicamente procesos intensivos de aplicación de diversas fuentes de conocimiento de orígenes y naturaleza diferente, que se corresponden con los distintos niveles de representación del discurso hablado:

- Fonético-acústico: relativo a las propiedades y características de los sonidos del habla y la determinación de las unidades elementales de éstos.
- Fonológico: la alteración de los sonidos en diferentes contextos como consecuencia de fenómenos de coarticulación, asimilación, etc.
- Léxico: el conocimiento léxico permite crear asociaciones entre secuencias de unidades fonéticas y palabras.
- Sintáctico: controla el proceso de concatenación de palabras (oraciones) que

son gramaticalmente correctas de acuerdo a un determinado modelo de lengua.

- Semántico: relativo al significado de las palabras y oraciones basadas en la manera en que las palabras se usan de manera más normal. El conocimiento semántico trata con el contenido de los mensajes (estructura interna) y puede desechar oraciones, incluso si son gramaticalmente correctas.
- Contextual: ayuda a la compresión de oraciones teniendo en cuenta el entorno de la tarea o diálogo en el que se ha producido.
- Prosódico: las características prosódicas (entonación, duración, intensidad) sirven también para modifican el significado de un mensaje. Incluso en algunas lenguas influye en el nivel fonológico (ej. chino).

El número de trabajos dedicados a estudiar diferentes aspectos del campo fonético-acústico, como puede ser la descripción espectral de grupos de sonidos es muy numeroso; no así los dedicados a la aplicación de este conocimiento al campo del reconocimiento automático de la voz.

La panoplia de estudios, que intentan descubrir las estructuras subyacentes tras el fenómeno de la comunicación, puede dividirse en dos grandes conjuntos: los dedicados al estudio de las vocales y los que centran en los rasgos que permiten clasificar a los sonidos consonánticos.

Dentro del primer grupo encontramos el trabajo clásico de Peterson y Barney [PET52]. En éste trabajo, uno de los más influyentes, se recogen conjuntos de medidas para los formantes F_1 al F_3 y para el tono fundamental (F_0). Con las medidas de los formantes F_1 y F_2 los autores elaboran el triángulo vocálico de las vocales inglesas; apreciándose solapamientos para distintas ocurrencias entre vocales adyacentes, aunque con este método existe un alto grado de acierto en su clasificación.

Aunque este estudio, ya marca la pauta acerca de los problemas de variabilidad en el habla entre diferentes locutores, presenta algunas deficiencias importantes [HIL95]: no se tiene información de las variantes dialectales, los resultados no se agrupan por género o edad y en cuanto al grupo de los niños no se conoce su edad ni el tamaño de la

población es suficiente.

Otros trabajos acerca de vocales, analizan la importancia de las transiciones vocálicas al principio y final de éstas [ASS96]. Si bien, las transiciones de los formantes no ayudan a la determinación de las vocales, son beneficiosas porque proporcionan al oyente una pista que ayuda al reconocimiento del habla en las regiones de transición entre vocal y consonante. Esta es una de las razones de la mayor importancia de los sonidos vocálicos en la inteligibilidad de los mensajes hablados [COL96].

Por último, dentro de este conjunto son también importantes los estudios acerca de los diptongos, como por ejemplo [AGU95], [NAB96] para la producción de diptongos del castellano en diversas situaciones de habla espontánea, o [SAM95] que estudia los diptongos del inglés australiano.

Dentro del grupo de trabajos dedicados a las consonantes, muchos de los que podemos encontrar se ocupan de la detección del punto de articulación para los sonidos oclusivos. Este caso es uno de más difíciles de reconocer, tanto para oyentes humanos como para los sistemas artificiales, aunque las capacidades de estos últimos están muy por detrás [MOR97].

Estos análisis se centran en determinar el punto de articulación de las consonantes oclusivas, a partir de la barra de explosión y la transición de formantes [BLU79], [BON96], [BON97], [ESP97], [NEA97]. La barra de explosión de forma aislada proporciona pistas invariantes e independientes del contexto para la identificación de las consonantes oclusivas, sin embargo, la posición y evolución de los formantes de la siguiente vocal son necesarios para el perfecto reconocimiento de las secuencias de voz.

Otros estudios [DJE95] [NIY98], se centran en la información proporcionada por la medida del VOT (*Voice Onset Time*), es decir, el tiempo que transcurre entre la explosión y la consolidación de la vocal posterior a una oclusiva.

Las investigaciones también se ocupan de los otros grupos consonáticos como por ejemplo:

- [SHA88] modelos auditivos aplicados a detección de rasgos fonéticos de vocales y fricativas sordas.
- [OHD96] estudia las nasales para niños y adultos.
- [ABD98] describe un sistema para el reconocimiento de consonantes fricativas del inglés americano a partir de tres medidas tomadas de bandas del espectro: pendiente espectral normalizada máxima, posición del pico más dominante y centro de gravedad espectral.

El volumen de trabajos dedicados a la aplicación de los conocimientos anteriores a sistemas de reconocimiento es mucho menor:

- En [LIU96] se describe la estructura de un reconocedor basado en la detección de una serie de marcas o guías, que servirían como rasgos distintivos. El algoritmo detecta las marcas producidas por cambios acústicos bruscos como son: cierres y aperturas en oclusivas, nasales, así como el inicio y fin de la vibración de las cuerdas vocales. Emplea diversos factores para calcular las marcas: cambios repentinos en la energía, diferentes niveles de resolución, duración de segmentos y restricciones articulatorias. Los resultados de identificación de las marcas están por encima del 90% de acierto para habla limpia. Sin embargo, el sistema reconocedor asociado a esas medidas de marcas está todavía en una fase inicial.

- [HAL97] presenta un modelado fonético-acústico empleando conjuntos heterogéneos de medidas. La idea es que diferentes medidas pueden describir mejor determinados grupos fonéticos que otros. Toma conjuntos de parámetros separados y adaptados para la detección de vocales, nasales, oclusivas y fricativas estimados sobre diferentes ventanas, incluyendo el tono fundamental, energía de frecuencias bajas y cruces por cero. La combinación de vectores se realiza teniendo en cuenta la probabilidad de que estemos ante un determinado sonido en función de los valores de cada vector de parámetros obtenidos. Comparado con un reconocedor de fonemas basado en coeficientes cepstrales y escalas de Mel obtiene un resultado ligeramente mejor.

Por otra parte y aunque encontramos diferentes estudios que intentan determinar la

importancia que tienen el tono fundamental en el proceso de percepción del habla, como por ejemplo [KEW96], que analiza el influjo del tono fundamental en la discriminación de las frecuencias de los formantes vocálicos, lo cierto es que su aplicación al campo del reconocimiento de voz ha sido prácticamente nula en el pasado. Sin embargo en la actualidad es un aspecto que empieza a cobrar interés [KOM97], [THO98].

El tipo de aproximación más usual en los sistemas de reconocimiento cuando se desea incorporar conocimiento fonético-acústico suele ser bien, a través de algoritmos de detección de formantes, o mediante el uso de parámetros acústicos.

a).- Algoritmos de detección de formantes. Los métodos de extracción de formantes son procedimientos de cálculo que partiendo de las representaciones espectrales obtenidas de una traza de voz dada, determinan la posición de los formantes presentes en la señal. Históricamente las primeras herramientas de análisis que se emplearon en los algoritmos de búsqueda de formantes de la voz, estaban basadas en los métodos de predicción lineal [ITA70] [MAR72] [MCC74].

Las motivaciones que podemos encontrar para el análisis de este tipo de estructuras [SNE93], [NIR94] [PLA95], son diversas y abarcan desde el desarrollo de la tecnología de síntesis de voz hasta su aplicación en el estudio de la fonética [BOB98]. Dentro del campo del reconocimiento automático del habla [WEL95], [WEL98a], su interés principal se centra en las áreas de clasificación y caracterización de fonemas [LAP94], [SCH95], así como en su aplicación directa a la fase de extracción de parámetros [HOL97].

El funcionamiento de los algoritmos de cálculo de formantes se basa esencialmente, en técnicas de programación dinámica que permiten ir evaluando los distintos puntos candidatos. Sin embargo y con independencia del método empleado, aparecen ciertos casos dudosos que hay resolver. Para ello se aplican diversas restricciones:

1. Los formantes presentan formaciones que se caracterizan por su continuidad y por la suavidad de sus transiciones al menos en periodos cortos de tiempo.

2. Las frecuencias de aparición de los formantes no son aleatorias. Para cada uno

de ellos se puede establecer una banda de frecuencias y determinar las relaciones posibles entre bandas para el conjunto de ellas.

3. Determinar los periodos sonoros, sordos y de silencio.

Una de las mayores dificultades que presenta la aplicación de información de formantes a los sistemas de reconocimiento, es la de la fiabilidad de los métodos de detección. En la práctica la utilización de este recurso debe ir acompañada de medidas de fiabilidad del proceso, como forma de garantizar el efecto positivo que la incorporación de esta fuente de información puede aportar a los vectores de rasgos acústicos [HOL97], [GAR98].

b).- Parámetros acústicos o AP (*Acoustic Parameters*) [BIT95]. Un parámetro acústico es una métrica, obtenida directamente del espectrograma de una realización de habla, que proporciona una estimación precisa de un rasgo acústico o articulatorio.

La hipótesis de trabajo consiste en considerar que los fonemas pueden clasificarse en sus diferentes clases por medio de un conjunto de rasgos fonéticos distintivos, y que a su vez, éstos pueden expresarse por medio de estimadores eficaces.

La principal motivación en el uso de los parámetros acústicos tiene su fundamento en su mayor la robustez frente a las diferencias presentes en el habla de varios locutores, en comparación con los sistemas de parametrización más usuales empleados en los sistemas de reconocimiento. La robustez, en este caso, se obtiene definiendo los parámetros acústicos de manera relacional a través de los dominios temporal y frecuencial.

Estos estimadores, basados en parte en algoritmos de detección de formantes, son además, una herramienta útil para el estudio de la fonética acústica y una forma intuitiva de revertir a su vez, los conocimientos de esta disciplina al campo del reconocimiento automático.

Algunos ejemplos de parámetros acústicos incluyen (Tabla 1.2):

a).- Medida de la energía en diferentes bandas de frecuencia y su relación con

medidas obtenidas para otras bandas.

b).- Estimadores de la rapidez de cambios en el espectro.

c).- Medidas de la posición relativa de frecuencias de formantes.

Rasgo	Parámetro acústico
Sonoro	Energía[100: 400Hz]
	Energía[0: 2kHz]
	Energía[0: 300Hz]: Energía[3.7: 7kHz]
	Energía[0: 688Hz]/ Energía[4kHz: 8kHz]
Vocálico	Energía[600: 2800Hz]
	Energía[2000: 3000Hz]
	Energía[0: 8kHz]
Consonante	Energía[0: 8kHz]
	Banda del F_1
	Banda del F_2
	Banda del F_3
	Banda del F_4
Palatal	Energía[(F_3: 187): (F_3+594)]/ E[0: 8kHz]
	Energía[(F_3: 781): (F_3+312)]/ E[0: (F_3: 781)]
Nasal	Valor absoluto[250Hz- F_1]
	Valor absoluto[2500Hz- F_2]
	Valor absoluto[3250Hz- F_3]
Punto de articulación de oclusivas	Energía[(F_3+31): (F_3+3250)]/ E[0: (F_3: 31)]
	Energía[(F_3-1750): F_3)/ E[0: 8kHz]
Grave/Agudo	Energía[0: 500Hz]- Energía[500: 1000Hz]
	F_1- F_0
Delantero/Trasero	Energía[800: 1900Hz]- Energía[2: 3kHz]
	F_2- F_1
Estridente	Energía[6: 8kHz]
	Energía[(F_3+562): (F_3+1125)]/ Energia_mínima[(F_3+562): (F_3+1125)]

Tabla 1.2. Ejemplos de parámetros acústicos [BIT97] [VAR97].

La motivación principal para incorporar conocimiento en los sistemas automáticos de reconocimiento es la de poder mejorar los modelos y algoritmos usados en diferentes pasos del proceso, al poder reducir el número de posibles interpretaciones con la introducción de restricciones en el análisis de las oraciones candidatas. Existen dos formas básicas de integrar el conocimiento lingüístico en los sistemas de reconocimiento:

1. Aproximación basada en los datos. En este tipo de enfoque el conocimiento se extrae de forma automática o semiautomática de grandes bases de datos de voz para almacenarse de manera implícita en modelos estocásticos (modelos ocultos de Markov y redes neutrales artificiales). Durante los últimos años mucho del esfuerzo investigador se ha concentrado en el estudio de estos modelos. Aun así dentro de este tipo de aproximación encontramos intentos de añadir conocimiento específico:

- La velocidad en la producción del habla y en consecuencia la duración de los sonidos, puede explotarse durante el propio reconocimiento o en una fase posterior. [MAR98], [PFA98].
- Algunos estudios [MAR94], [TEB95], [MAR97] cuestionan algunos postulados poco realistas de los HMM; en particular la asunción de independencia que implica que todas las observaciones dependen solamente del estado que las genera.
- Algunas de las variaciones fonéticas pueden ser capturadas mediante modelados que tengan en cuenta los contextos de los sonidos [LEE90c].

Otra posible solución es utilizar redes neuronales para computar conjuntos de símbolos adicionales, que puedan entregarse como observaciones transformadas para los HMM [BEN90]. Más específicamente, la red genera grados de certeza para los rasgos básicos de los sonidos como sonoridad, fricación y oclusión/silencio. Una mejora de este método se consigue a través de la optimización global de la red neuronal y los modelos ocultos de Markov [BEN92]. En este caso se emplea el gradiente del criterio de optimización de los HMM con respecto a la transformación de las observaciones, como forma de estimar los pesos de la red.

Otra solución con red neuronal consiste en emplearla como forma de etiquetar las observaciones de un sistema discreto de modelos ocultos de Markov [LEC94]. Aquí en vez de utilizarla como estimador de probabilidad, lo que se realiza es un proceso de cuantificación vectorial donde cada clase representará una etiqueta fonética

2. Aproximación basada en el conocimiento. El objetivo aquí es modelar y usar el conocimiento lingüístico de forma explícita, a través de técnicas de razonamiento

heredadas de los campos de la inteligencia artificial e ingeniería del conocimiento. El conocimiento se obtiene de expertos humanos y el gran reto supone implementar de manera eficaz la forma en que las fuentes de conocimiento cooperan entre sí [HAT85]. Los modelos y estrategias de planificación y coordinación de las citadas fuentes pueden clasificarse en dos grandes grupos:

a).- Jerárquicos, donde cada fuente de conocimiento se comunica solamente con sus niveles adyacentes, ya sea, de arriba a abajo o viceversa.

b).- Modulares, donde la comunicación entre las fuentes independientes entre si, se realiza a través de elementos comunes como pueden ser mecanismos de paso de mensajes o modelos de pizarra [GON88].

La forma de representar el conocimiento adquirido del experto suele ser mediante reglas de producción en la forma condición-conclusión. Este tipo de mecanismo se ha empleado sobre todo en el proceso de decodificación acústico-fonética [GRE86], [ALL87]. La motivación es la existencia de una considerable cantidad de material con conocimiento explícito de las propiedades acústicas de las unidades fonéticas obtenidas a través de experimentos de lectura de espectrogramas. El sistema APHODEX (*Acoustic PHOnetic Decoding EXpert*) [CAR87] es un buen ejemplo de este último punto. El objetivo de este sistema era determinar los mecanismos subyacentes del proceso de decodificación fonética y el medio era el estudio de los métodos de razonamiento de un experto fonetista en la tarea de la interpretación de espectrogramas.

En concreto un experto emplea una estrategia para decodificar una oración que consta de dos fases: una exploración global del conjunto de la realización y un análisis de izquierda a derecha del espectrograma.

Durante la primera etapa, el experto trata de determinar los umbrales presentes en los dominios espectral y temporal para pasar a continuación a la detección de las zonas vocálicas y calcular la duración vocálica media. Este parámetro proporciona una idea del número de segmentos fonéticos presentes en la realización y desempeña un papel importante en la decodificación de las clases vocálicas.

Durante la etapa de análisis, el experto de forma iterativa busca aquellos rasgos

acústicos no ambiguos, emite una lista de hipótesis fonéticas y por último busca pistas que refuercen o invaliden las anteriores hipótesis.

A lo largo del proceso el experto emplea por una parte su experiencia para extraer la información relevante del espectrograma y aplica su conocimiento para realizar razonamientos específicos sobre la información anterior.

A pesar de los resultados positivos obtenidos con este tipo de aproximación encontramos dos problemas principales: el primero es la falta de certeza en los análisis producidos por las herramientas automáticas de detección acústica (detectores de formantes, analizadores de bandas de explosión, etc.). El segundo es la posible ausencia de determinadas claves acústicas y el diferente grado de discriminación que determinadas claves acústicas poseen en realizaciones concretas.

Con objeto de sobrellevar estas dificultades, muchos sistemas usan mecanismos de combinación de las puntuaciones asignadas a las diferentes soluciones parciales. Sin embargo, el ajuste del mecanismo de combinación es complicado y además este sistema no impide que hipótesis incoherentes resulten ser las que obtienen mejores puntuaciones.

Estas dificultades han llevado a que se construyan sistemas cuyo objetivo es preservar la coherencia de los razonamientos que se llevan acabo durante el proceso de reconocimiento al mismo tiempo que se comprende, explica y eventualmente se cuestiona cada decisión. Para ello se suelen emplear los trabajos previos del campo de la inteligencia artificial acerca del razonamiento hipotético [BON92], generación de hipótesis (*abduction*) y revisión de éstas (*truth maintance*) [DEM84].

2. REVISIÓN DE LOS ASPECTOS GENERALES DE FONÉTICA ACÚSTICA

2.1. ASPECTOS GENERALES DE LA PRODUCCIÓN Y PERCEPCIÓN DE LA VOZ

2.1.1. Fonética y Fonología

2.1.1.1. El esquema de la comunicación

La comunicación se puede definir como la transmisión de un lugar a otro de una determinada información. El sistema de comunicación está constituido por todos los elementos que intervienen en esta transmisión (Figura 2.1):

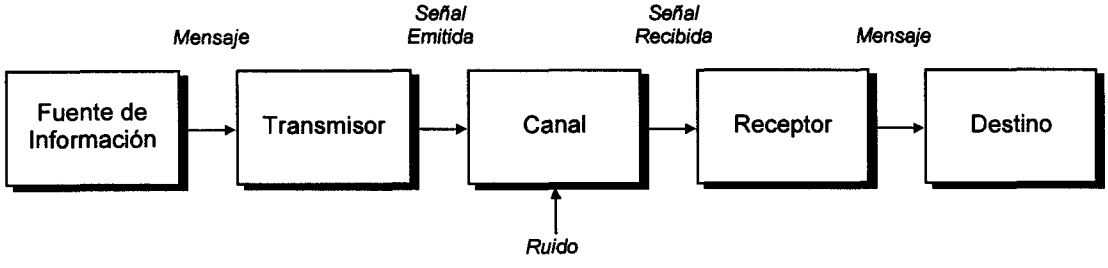

Figura 2.1. Esquema de la comunicación.

1. Emisor o fuente de información. Es el origen de lo que se va a comunicar. Este papel puede ser desempeñado bien por el hombre o bien por un sintetizador de voz. El emisor debe seleccionar de entre un conjunto de signos, convenidos previamente con el objeto de comunicar una información, para constituir un mensaje.

2. Destino. El destino es el lugar donde se recibe el mensaje. Puede ser el hombre o también una máquina.

3. Canal de transmisión. Es la unión espacial o temporal existente entre la fuente de información y el destino.

4. Transmisor. Realiza el paso de información entre el emisor y el canal. La operación que realiza el transmisor se denomina codificación, es decir, la conversión del mensaje estructurado en un código. Un código es un conjunto de reglas no ambiguas, previamente convenidas. Un ejemplo de transmisor puede

ser el aparato vocal (fonador) del hombre.

5. Receptor. Es la unión entre el canal y el destino. Realiza una nueva transformación en la información, es decir, realiza una operación inversa a la del transmisor: la decodificación del mensaje. En la comunicación oral, un ejemplo de receptor es el oído.

El lenguaje humano es un sistema de comunicación que utiliza determinados signos previamente convenidos; son los denominados signos lingüísticos. Un signo lingüístico está formado por un significante, que sería la suma de unos elementos fónicos (sonidos o fonemas según el nivel de análisis) y por un significado, que sería la idea o el concepto que nosotros tenemos acerca de ese signo. El significante en el habla representa un número infinito de realizaciones articulatorias, pero en la lengua, sin embargo, este número es finito.

Así, el estudio de los sonidos del lenguaje se realiza desde dos puntos de vista: lengua y habla, aunque fuertemente interrelacionados [QUI93], [ALA91]:

- Desde el punto de vista de su función en el sistema de la comunicación lingüística ⇒ ***Fonología***.
- Desde el punto de vista de su producción, de su constitución acústica y de su percepción ⇒ ***Fonética***.

En la comunicación oral, una sucesión de sonidos, conformados acústicamente, llega al oído del interlocutor, que reacciona ante ellos porque son los portadores de la información. Es decir, lo inmediato es lo acústico, que es lo que siempre percibimos, aunque el emisor, o creador de la onda sonora por medio de sus órganos articulatorios, no esté presente.

La fonética acústica estudia los mecanismos por los cuales el oyente es capaz de identificar las unidades de comunicación que constituyen el mensaje hablado. Tradicionalmente se ha estudiado la fonética desde el punto de vista genético o articulatorio, es decir, desde el punto de vista del hablante sin tener en cuenta el lado acústico, el del oyente. No obstante la fonética articulatoria es el origen de la acústica;

sin ella no podríamos explicar la formación de cada sonido ni de los hábitos motores articulatorios, que tan importante papel desempeñan en la identificación de la onda acústica percibida.

2.1.1.2. Alfabetos fonéticos

Existen diferentes alfabetos fonéticos para la tarea de transcribir los sonidos de una lengua. Algunos son propios de los desarrolladores de aplicaciones de tratamiento de la señal de voz. Los más conocidos son:

1. Alfabeto Fonético Internacional (AFI) o International Phonetic Alphabet (IPA). Con 100 años de historia y cada vez más extendido. Será el que empleemos en este trabajo.
2. Revista de Filología española, RFE. El más empleado en los países de habla española.
3. ARPAbet. Fue desarrollado en los años 70 para soportar los proyectos ARPA.
4. TIMIT. Variante del anterior usado en la base de datos del mismo nombre.
5. .Worldbet. Desarrollado por Jim Hieronymous (AT&T) para manejar diferentes lenguas con un único sistema ASCII.

2.1.1.3. Unidades fonológicas y fonéticas

Los sonidos de una lengua pueden sufrir modificaciones sin que varíe en nada su valor significativo. Este tipo de modificación o rasgo será de tipo **no funcional** o **no pertinente**. Un ejemplo puede se la diferencia fricación/oclusión en la realización de *el bote* [el βóte] y *un bote* [úm bóte], sin que cambie el significado de la palabra *bote*.

Sin embargo, existe en una lengua todo un gran conjunto de sonidos en los que al variar un rasgo determinado, cambia radicalmente su significado; éste será un **rasgo pertinente** o **funcional**. Ejemplo rasgo sorda/sonora en la oclusiva bilabial de las palabras *poca* [póka] y *boca* [bóka].

Fonema es la unidad fonológica más pequeña en que puede dividirse un conjunto fónico. Ejemplo: /páso/ paso, está formada por cuatro fonemas, ya que es el máximo de

unidades mínimas en que puede ser dividida es /p/+ /a/+ /s/+ /o/. Se representan mediante barras oblicuas. La Tabla 2.1 muestra el conjunto de fonemas del español.

Consonantes	Bilabial		Labiodental		Linguodental		Linguointerdental	
	Sorda	Sonora	Sorda	Sonora	Sorda	Sonora	Sorda	Sonora
Oclusiva	p	b			t	d		
Fricativa			f				ɵ	
Africada								
Nasal		m						
Lateral								
Vibrante simple								
Vibrante múltiple								

Consonantes	Linguoalveolar		Linguopalatal		Linguovelar	
	Sorda	Sonora	Sorda	Sonora	Sorda	Sonora
Oclusiva					k	g
Fricativa	s			ʃ̂	x	
Africada			c			
Nasal		n		ɲ		
Lateral		l		ʎ		
Vibrante simple		r				
Vibrante múltiple		r̄				

Vocales	Anterior	Central	Posterior
Cerrada	i		u
Media	e		o
Abierta		a	

Tabla 2.1. Cuadro de los fonemas de español [QUI96].

Alófono es cada una de las realizaciones fonéticas que puede tener un fonema, de acuerdo con el contexto en el que se halle situado. Se representan siempre entre corchetes. Por ejemplo el inglés tiene dos alófonos del fonema /p/. En posición inicial de palabra se pronuncia [p^h] como en [p^hil] *pill*, es decir seguido de una aspiración, pero como [p] en [spil] *spill*. Las distintas realizaciones de un fonema dependen del contorno en que se encuentren situadas, pero no implica un cambio de significación.

Hablamos de **distribución complementaria** cuando los alófonos de un determinado fonema aparecen en unas posiciones concretas y no en otras. Por ejemplo, en español, el fonema /b/ conoce un alófono oclusivo [b] que se realiza como tal después de pausa y después de consonante nasal, y otro alófono fricativo [β] en los demás contornos. Es decir, que en un contorno donde se da un alófono no aparece normalmente otro. Por el contrario, la distribución libre no implica que en una posición determinada se encuentre normalmente un sólo alófono. Ejemplo: el archifonema /R/ en

español al final de palabra.

La relación que existe entre dos fonemas conmutables, es decir, sustituyendo uno por el otro, recibe el nombre de **oposición**. La conmutación solamente es necesario realizarla para aquellos fonemas que presente características similares. En inglés, por ejemplo, /sip/ sip, se opone a /zip/ zip, porque la primera, /s/, es sorda y la segunda /z/, sonora; constituyendo por tanto /s/ y /z/ una oposición.

Archifonema o neutralización es el resultado de una pérdida de la función distintiva para dos fonemas en ciertas posiciones. Se suele representar por una letra mayúscula entre barras. Un ejemplo para el español es el caso de la vibrante simple y la vibrante múltiple. En posición interior de palabra implica una diferenciación significativa pero no así en posición final de sílaba.

Consonantes	Bilabial		Labiodental		Linguodental		Linguointerdental	
	Sorda	Sonora	Sorda	Sonora	Sorda	Sonora	Sorda	Sonora
Oclusiva	p	b			t	d		
Fricativa		β	f				ɵ	ð
Africada								
Nasal		m		ɱ		n̪		n̪
Lateral						l̪		l̪
Vibrante simple								
Vibrante múltiple								

Consonantes	Linguoalveolar		Linguopalatal		Linguovelar	
	Sorda	Sonora	Sorda	Sonora	Sorda	Sonora
Oclusiva					k	g
Fricativa	s	ɹ s̬		ʝ̂	x	ɣ
Africada			c	ɟ		
Nasal		n		ɳ ɲ		ŋ
Lateral		l		ʎ		
Vibrante simple		r				
Vibrante múltiple		r̄				

Vocales	Anterior	Central	Posterior
Semiconsonante	j		ɯ
Semivocal	i̯		ʊ
Cerrada	i		u
Media	e		o
Abierta		a	

Tabla 2.2. Cuadro de los sonidos del español [QUI96].

Consonantes	Bilabial		Labiodental		Linguointerdental		Linguoalveolar	
	Sorda	Sonora	Sorda	Sonora	Sorda	Sonora	Sorda	Sonora
Oclusiva	p	b					t	d
Africada								
Nasal		m						n
Lateral								l
Vibrante simple								
Vibrante múltiple								
Fricativa			f	v	θ	ð	s	z r

Consonantes	Linguopalatal		Linguovelar		Glotal	
	Sorda	Sonora	Sorda	Sonora	Sorda	Sonora
Oclusiva			k	g		
Africada	c	ɟ				
Nasal				ŋ		
Lateral				ʟ		
Vibrante simple						
Vibrante múltiple						
Fricativa	ʃ	ʒ			h	

Semiconsonantes	Bilabial		Labiodental		Linguointerdental		Linguoalveolar	
	Sorda	Sonora	Sorda	Sonora	Sorda	Sonora	Sorda	Sonora
Semiconsonante		ɯ						

Semiconsonantes	Linguopalatal		Linguovelar		Glotal	
	Sorda	Sonora	Sorda	Sonora	Sorda	Sonora
Semiconsonante		j				

Vocales	Anterior	Central	Posterior
Cerrada	i:		u:
	i		u
Medio-cerrada	e	ə:	o
Medio-abierta	ɛ	ə	ɔ:
Abierta	æ	ʌ	ɔ
		a	ɑ

Tabla 2.3. Cuadro de los sonidos del inglés [QUI96].

La Tabla 2.2 y la Tabla 2.3 contienen los cuadros con los sonidos del español y del inglés, respectivamente. Aunque los dos grupos presentan sonidos comunes, encontramos bastantes elementos que solamente están en alguno de ellos. Este hecho explica el porqué las realizaciones producidas por locutores no nativos presentan una mayor variabilidad. En muchos casos, un hablante de una segunda lengua tiende a asimilar los sonidos no presentes en su idioma convirtiéndolos en alguno de los existentes en el suyo propio. Como veremos más adelante, este fenómeno afecta de manera notable al proceso de reconocimiento de voz.

2.1.2. Producción del sonido articulado

Durante la fonación interviene un conjunto de órganos que pueden ser clasificados en tres grupos bastante bien diferenciados:

1. Cavidades infraglóticas. Formadas por los órganos propios de la respiración: pulmones, bronquios y traquea. Los pulmones son los que presentan un papel más relevante; son los proveedores de la cantidad de aire suficiente para que el acto de la fonación sea realizable.

2. Cavidad glótica. Se halla delimitada por una serie de cartílagos y músculos situados a la altura de lo que en los hombres se conoce como "nuez". El sistema de cartílagos y músculos, en una determinada posición permite la libre circulación de gas (posición respiratoria), mientras que en otra determinada posición cierran total o parcialmente el flujo de aire procedente de los pulmones (posición fonadora) mediante un sistema de membranas adherido a los cartílagos.

Dichas membranas de tipo muscular reciben el nombre de cuerdas vocales. En posición de cierre, producen una retención de flujo procedente de los pulmones, con el consiguiente aumento de presión. Cuando la presión supera la fuerza elástica de las cuerdas, éstas se separan ligeramente y permiten el paso de una pequeña descarga de fluido, con lo cual se reduce la presión en la zona de paso, y la fuerza de recuperación de las cuerdas vuelve a producir el cierre de la abertura. La presión se vuelve entonces a incrementar, y fuerza una nueva descarga de flujo, y así sucesivamente. Estas descargas de flujo aéreo son conocidas como pulsos glotales, se producen a razón de varios cientos de veces por segundo y constituyen la base de la ***fonación sonora***.

En la formación del *sonido vocálico*, las cuerdas vocales están más tensas, por consiguiente los golpes vibratorios son más fuertes y la frecuencia también es mayor. El grado de abertura de la glotis es mínimo y, por lo tanto, también lo es el gasto de aire. En la formación del *sonido consonántico* sonoro, las cuerdas vocales están menos tensas y el grado de abertura de la glotis es mayor con el consiguiente gasto de aire

3. Cavidades supraglóticas. Se encuentran por encima de la cavidad glótica, y

están constituidas fundamentalmente por la cavidad faríngea, hasta llegar a la raíz de la lengua y la úvula, en cuyo punto se divide en dos nuevas cavidades: la cavidad nasal y la cavidad oral. El conjunto de cavidades se muestra en la Figura 2.2.

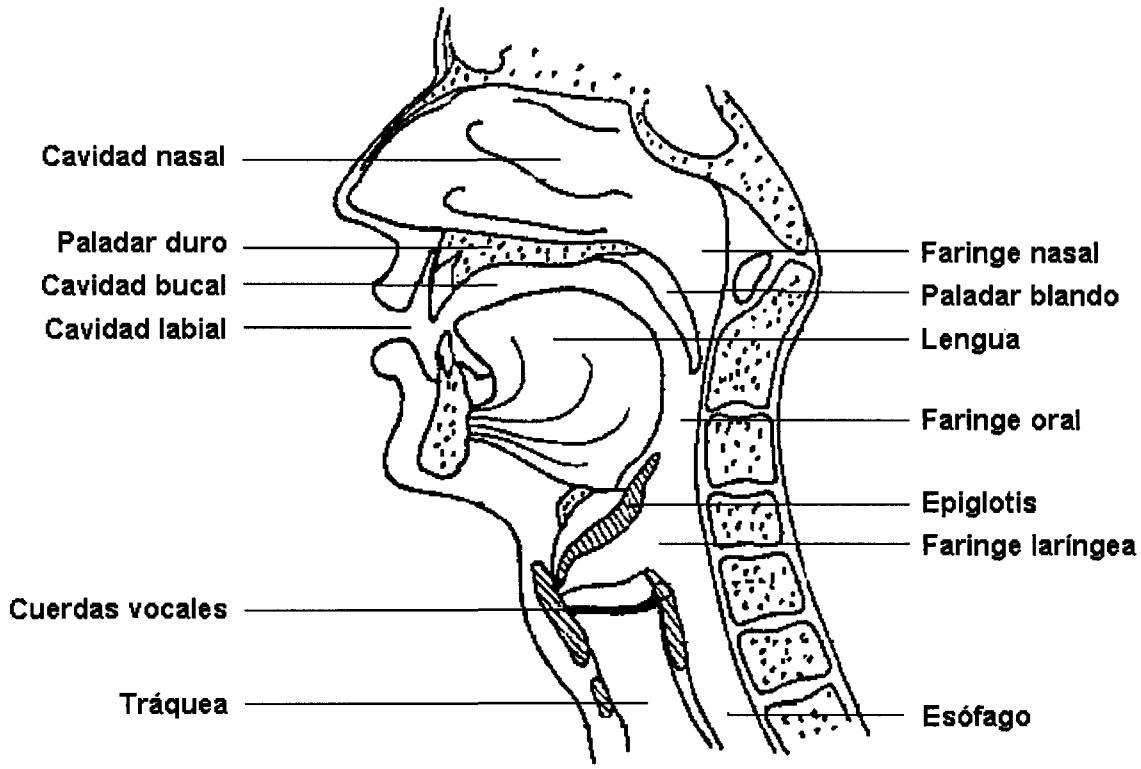

Figura 2.2. Corte vertical de los órganos fonadores.

El tabique que inicia la división entre las cavidades oral y nasal se denomina paladar blando o velo, y por aproximación a la pared posterior de la faringe puede cerrar el acceso a la cavidad nasal. Siguiendo hacia el exterior, encontramos la pared superior de la cavidad oral, ya reforzada por estructura ósea, que se denomina paladar duro, presentando una forma abovedada. Su terminación en un arco separado de la encía y dientes superiores, recibe el nombre de arco alveolar. La parte superior de la cavidad oral termina en la cadena de dientes y por fuera de la misma en el labio superior.

Por lo que respecta a la parte inferior de la cavidad oral, arrancando de la raíz de la lengua, encontramos la zona medial de la misma, correspondiendo al comienzo del paladar duro, y más adelante, la zona apical o punta de la lengua. Las zonas laterales, se extienden a ambos lados de la zona apical en unos 2 o 3 cm, también sobre la lengua. La parte inferior termina en la cadena de dientes inferior y en el labio inferior.

La cavidad nasal consiste en una estructura recubierta por paredes mucosas, permaneciendo su forma relativamente inalterada durante la articulación de sonidos.

La codificación que se realiza sobre los sonidos sonoros y sordos tiene lugar debido a los puntos en los que se producen estrechamientos voluntarios de las cavidades supraglóticas. Estos puntos se conocen como puntos de articulación y son controlados por los órganos móviles (velo, raíz, medio, ápice, laterales, corona dental inferior, labios) contra los órganos inmóviles (pared faríngea, paladar duro, alvéolos, corona dental superior). Los diferentes puntos de articulación tienen lugar por medio de los adecuados estrechamientos u oclusiones del tracto vocal entre un órgano fijo y uno móvil, dentro de aquellos ubicados en una misma zona, y que por ello pueden trabar contacto.

2.1.2.1. Modelos articulatorios

A lo largo del tiempo se han propuesto numerosos modelos del sistema de producción de la voz humana. Los más comunes antes de la generalización de los modelos digitales, han sido los basados en modelos eléctricos analógicos [FAN73]. Estos modelos fueron desarrollados para diversos propósitos, desde investigación básica hasta el diseño de sistemas articulatorios dedicados a la síntesis de la voz.

Aunque pocos estudios tratan de los aspectos articulatorios del reconocimiento de la voz [DEN94a], lo cierto es que merece la pena repasar los fundamentos de dichos modelos. El proceso de la producción del habla puede modelarse a grandes líneas tal y como se indica en la Figura 2.3.

Una aproximación razonable consiste en modelar el aparato fonador como un tubo cilíndrico, no uniforme y compuesto por diferentes secciones [DEL93]. Destacan sobremanera las configuraciones de 2 y 4 tubos para el estudio de la producción de las vocales [FAN73].

2.1.2.2. Producción del habla en condiciones de ruido

Un fenómeno conocido desde hace tiempo es que una fuerte presencia de ruido de fondo hace que el hablante incremente su esfuerzo vocal. A este fenómeno se le conoce

como efecto Lombard o reflejo Lombard (E. Lombard, 1911) y ha sido investigado en diferentes formas. Sin embargo dentro del ámbito del estudio del reconocimiento de la voz sus consecuencias han sido largamente ignoradas hasta hace unos años con la introducción de sistemas reconocedores para usos militares (cabinas de pilotos).

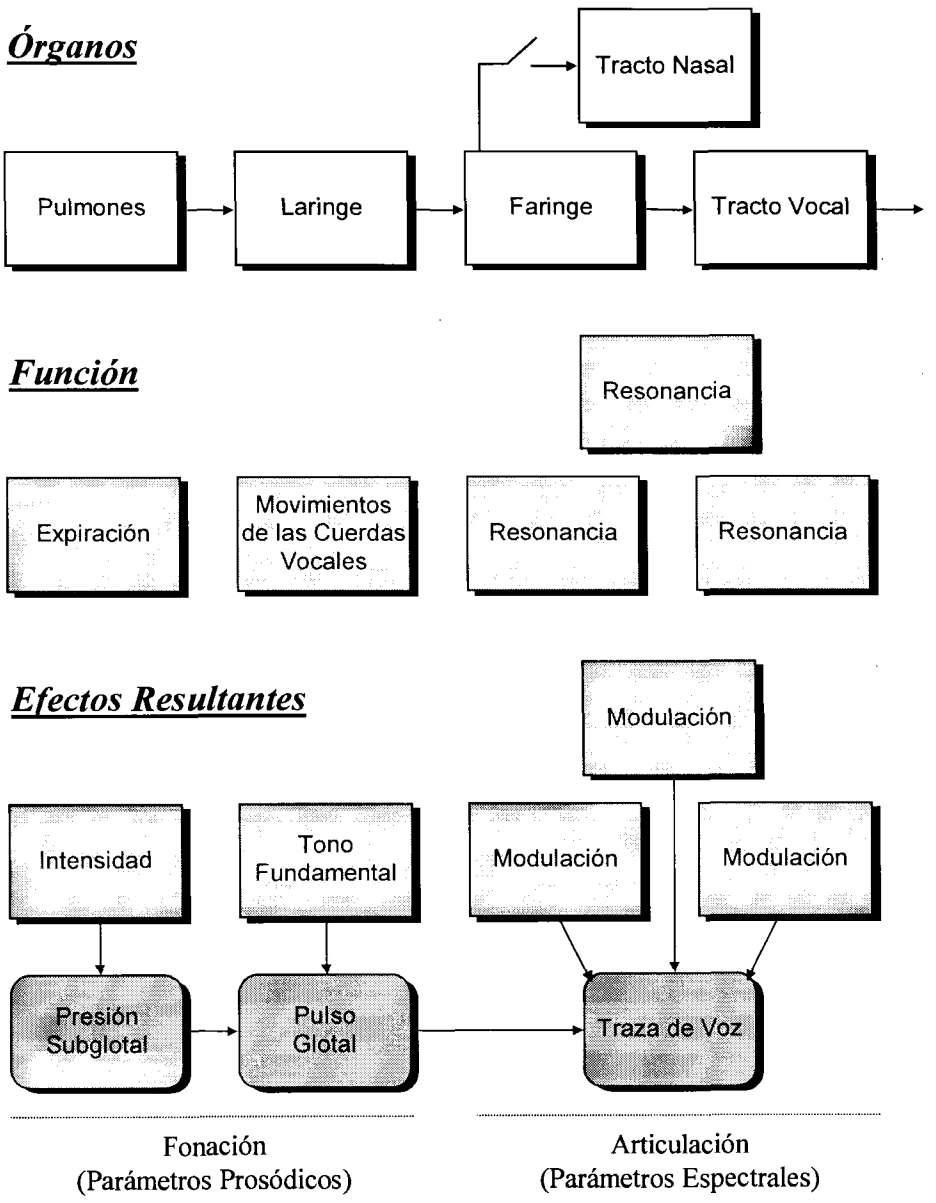

Figura 2.3. Diagrama funcional del aparato fonador [JUN96].

Desde el punto de vista de la producción de la voz, el efecto Lombard se corresponde con una amplificación de los movimientos articulatorios, a la vez que se mantienen las distinciones presentes en el habla normal. Esto resulta en un sustancial incremento en la intensidad glotal que a su vez produce un aumento en el flujo aéreo a

través de la cavidad oral. Se hacen necesarias, por tanto, compensaciones articulatorias (tales como desplazamientos de la lengua hacia abajo o articulaciones más abiertas) con objeto de evitar turbulencias a través de los puntos de constriccion del tracto vocal que pudieran deteriorar la cualidad vocálica de forma inaceptable [SCH89]. De esta forma la producción de la voz en entornos ruidosos no puede describirse como una amplificación lineal de la articulación normal, sino que se corresponde con una reorganización de los movimientos articulatorios [JUN96].

Aplicados al campo del reconocimiento automático, existen diferentes estudios que se fijan en las diferencias entre habla normal y Lombard [JUN89] [JUN93] [CAS95] [JUN95]. Esencialmente, las diferencias son cambios significativos en la estructura espacio-temporal del habla. Una enumeración de los principales cambios incluye:

- ✓ Incremento en la frecuencia fundamental F_0.
- ✓ Incremento en la amplitud.
- ✓ Incremento en la duración de las vocales.
- ✓ Incremento en el *spectral tilt.*
- ✓ Desplazamiento de los centros de frecuencias para F_1 (sobre todo) y F_2.

Aún así, se aprecia una variabilidad entre hablantes que dificulta la creación de modelos de habla Lombard unificados, que puedan emplearse en sistemas de reconocimiento.

2.1.3. Clasificación de los sonidos del lenguaje

Realizaremos esta clasificación basándonos en varios criterios:

2.1.3.1. Por la acción de las cuerdas vocales

Según vibren o no las cuerdas vocales, todo el material fónico se clasifica en *sonidos articulados sonoros* y *sonidos articulados sordos*. En español, todos los sonidos sordos son consonantes. A una consonante normalmente sonora, que por asimilación pierde parte de la sonoridad, se le llama *ensordecida*; y, a la inversa, una consonante sorda puede adquirir cierto grado de sonoridad, que la convierte en *sonorizada*.

2.1.3.2. Vocales y consonantes

Esta clasificación, que tiene su origen en las descripciones gramaticales griegas, ha prevalecido hasta nuestros días. No obstante la clasificación entre vocal y consonante tiene un fundamento fisiológico, que se manifiesta claramente en los resultados acústicos, y que está apoyado por los siguientes hechos:

1) Son distintos los músculos que intervienen en la formación de unos y otros. Para la producción de las vocales actúan los músculos depresores, que infieren en el maxilar inferior un movimiento de abertura. Estos músculos están también ligados también con los linguales De este modo, al descender el maxilar, la lengua también lo hace. Por ello, se han considerado los sonidos articulados vocálicos como los más abiertos.

Para la emisión de las consonantes actúan los músculos elevadores. Estos músculos tienden hacia la cerrazón del maxilar y como también están en contacto con los linguales, proyectan la lengua hacia el paladar (duro o blando). Por ello, los sonidos consonánticos se consideran como los más cerrados.

2). Por la acción de las cuerdas vocales. Para la producción del sonido articulado vocálico las cuerdas vocales presentan una mayor tensión, y un mayor número de vibraciones por unidad de tiempo, por lo que el tono fundamental de un sonido vocálico es siempre más alto que el de un sonido consonántico.

2.1.3.3. Por la acción del velo del paladar

Por la acción del velo del paladar los sonidos se clasifican en *orales* y *nasales*.

1) Los sonidos *orales* o *bucales* se producen cuando el velo del paladar está separado de la pared faríngea, y entonces el aire sale solamente a través de la cavidad bucal. Por ejemplo, la realización del fonema /m/ de /áma/ *ama*.

2) *Los sonidos nasales* se producen cuando el velo del paladar está separado de la pared faríngea, estando, por lo tanto, abierto el conducto nasal. Pueden ocurrir dos casos:

El primero que el velo del paladar se encuentre abierto, y la cavidad bucal totalmente cerrada.

El segundo caso se produce cuando el velo del paladar está separado de la pared faríngea, pero al mismo tiempo, el conducto oral también lo está. Este es el caso de vocales nasalizadas. En este caso es mejor el dar a este tipo de sonidos el nombre de oronasales.

2.1.3.4. Por el modo de articulación

Se da el nombre de modo de articulación a la posición que adoptan los órganos articulatorios en cuenta a su grado de abertura o cerrazón. De esta manera, en un sentido general, encontraremos *sonidos abiertos*, como las vocales; *medio cerrados* o *continuos*, como algunas consonantes: [s, f, x], etc.; y *cerrados* o *interruptos*: [p, t, k, b], etc.

Con arreglo a este concepto, podemos dividir los sonidos articulados del siguiente modo:

Las vocales, se dividen en:

1. *Altas*, *cerradas*, *extremas* o de *pequeña abertura*, si la lengua se aproxima hasta un máximo permisible para la articulación vocálica, bien el paladar duro o al paladar blando, como por ejemplo [i, u]

2. *Medias* o de *abertura media*, si la lengua se separa más de la bóveda de la

cavidad bucal, como por ejemplo [e, o].

3. *Bajas, abiertas* o de *gran abertura,* si la lengua se separa aún más de la bóveda palatal y ocupa un límite máximo de alejamiento, como por ejemplo [a].

Las consonantes, en:

1. *Oclusivas*, cuando hay un cierre completo de los órganos articulatorios. También se llaman *explosivas* o *momentáneas*. Las oclusivas en algunas lenguas pueden ser *aspiradas*. Esto ocurre cuando la explosión se realiza con la glotis abierta, dando lugar a que se oiga netamente el ruido de la espiración entre la explosión de la consonante y el comienzo de la vocal siguiente. Las que se pronuncian con la glotis cerrada son *no aspiradas* o *puras*, como las españolas [p, t, k].

2. *Fricativas*, cuando el sonido se forma a causa de un estrechamiento de dos órganos articulatorios, sin que éstos lleguen nunca a juntarse. También se denominan espirantes, constrictivas o continuas. Como por ejemplo: [f, θ, s, x], etc.

3. *Africadas* o *semioclusivas*, cuando al cierre completo de dos órganos articulatorios sucede una pequeña abertura por donde se desliza el aire contenido en el primer momento de cierre, percibiéndose claramente la característica de fricación. Por tanto podemos decir que una consonante africada consta de dos momentos: un primer momento de oclusión seguido inmediatamente de otro de fricación. En español, solamente existe un fonema africado [c]. En otras lenguas, son varias las consonantes susceptibles de ser africadas.

4. *Nasales*, cuando la cavidad bucal está cerrada y el pasaje nasal abierto. Así por ejemplo en español, desde el punto de vista fonológico, encontramos: [m, n, ɲ].

5. *Líquidas*. Forman un grupo especial que comprende:

a).- *Laterales*, en cuya emisión el aire sale por un lado, o por los dos, de la cavidad bucal [l, λ],

b).- *Vibrantes*, cuya característica es una o varias vibraciones del ápice de la

lengua: [r, r̄].

2.1.3.5. Por el lugar de articulación

Las vocales se clasifican en:

1. *Anteriores o palatales*, cuando la lengua ocupa una posición articulatoria en la región delantera de la cavidad bucal, esto es, en la zona que se halla cubierta por el paladar duro. Dentro de esta serie se encuentran las vocales [i, e].
2. *Posteriores*, si en vez de ser la parte predorsal de la lengua la que se aproxima al paladar duro, como en el caso anterior, es el postdorso el que se acerca a la región posterior de la cavidad bucal, esto es, al velo del paladar o paladar blando. Ejemplos: [u, o].
3. *Centrales*, si, por último, el dorso de la lengua se encuentra en una región cubierta por el mediopaladar. En el sistema fonológico del español solamente se conoce una: [a].

Las consonantes, según los órganos que actúen, así como la zona donde se inician éstos, se pueden clasificar en:

1. Bilabiales: [p, b, m, β]
2. Labiodentales: [f, ɱ]
3. Linguodentales o dentales: [t, d, n̪, l̪]
4. Linguointerdentales o interdentales: [ð, θ, n̪, l̪]
5. Linguoalveolares o alveolares: [s, n, r, r̄, l]
6. Linguopalatales o palatales: [c, ɲ, ʎ̟, λ, ĵ]
7. Linguovelares o velares: [k, g, x, ɣ, ŋ]

2.1.4. Rasgos distintivos intrínsecos

Los rasgos distintivos intrínsecos [JAK63], [FAN73] son manifestaciones propias de los fonemas, como segmentos constitutivos y autónomos de la cadena hablada. La presencia o ausencia de un rasgo intrínseco puede cambiar, si es pertinente, la naturaleza de un fonema. Se dividen en dos categorías: los *rasgos de sonoridad* y los *rasgos de tonalidad.*

2.1.4.1. Rasgos de sonoridad

Utilizan principalmente la cantidad y/o la concentración de energía acústica en cada momento de la producción de un fonema. Los rasgos de sonoridad son:

1. **Vocálico/no vocálico**.- Desde el punto de vista acústico, el rasgo vocálico se caracteriza por la presencia de una estructura formántica netamente definida. En el rasgo no vocálico, está ausente esa estructura formántica.

2. **Consonántico/no consonántico**.- Acústicamente, el rasgo consonántico se muestra bajo el efecto de una disminución de la energía total y por la presencia de zonas de no resonancia en su espectro, mientras que el rasgo no consonántico se caracteriza por el efecto contrario.

3. **Compacto o denso/difuso**.- Desde el punto de vista acústico, el rasgo compacto se caracteriza por una concentración más elevada de energía en una zona relativamente estrecha, central de su espectro, acompañada de un aumento de la cantidad total de energía y su expansión en el tiempo. Por el contrario, el rasgo difuso se caracteriza por una concentración más reducida de energía en la zona central del espectro, acompañada de una disminución de la cantidad total de energía y su expansión en el tiempo.

4. **Tenso/laxo o flojo**.- En contraposición con los fonemas flojos, los correspondientes fonemas tensos manifiestan un intervalo de sonido más largo y una energía más amplia. En una vocal tensa la desviación de sus formantes respecto a una posición neutra es mayor que en la correspondiente vocal laxa. Esta vocal se encuentra

más cerca del centro del triángulo vocálico que la vocal tensa. En las consonantes, la tensión se manifiesta, en primer lugar, por la duración de su periodo sonoro, y en las explosivas, además, por la gran fuerza de explosión.

5. **Sonoro/sordo.**- El rasgo de sonoridad se caracteriza acústicamente por la superposición de una fuente armónica sonora que se refleja en el espectrograma como un formante de muy baja frecuencia, situado, lógicamente, en su parte inferior. El rasgo de sordez se manifiesta por la ausencia de ese formante. En el sistema fonológico español, este rasgo y el rasgo tenso/laxo son redundantes, no así en inglés, alemán, francés, etc.

6. **Nasal/oral.**- Acústicamente, el rasgo de nasalidad se manifiesta en el espectro de las vocales a través de una reducción en la intensidad del primer formante F_1, y en las consonantes, por la aparición de zonas de formantes en unas determinadas frecuencias, siendo lo contrario para el rasgo de oralidad.

7. **Interrupto/continuo.**- El rasgo interrupto se caracteriza acústicamente por un momento de silencio, por lo menos en las bandas de frecuencia situadas por encima de las vibraciones de las cuerdas vocales, al que sigue una barra perpendicular que es el resultado de la explosión (si es sonora aparecerá en su parte inferior su barra de sonoridad, solamente). El rasgo continuo se caracteriza por la presencia de frecuencias, ya sean armónicas o inarmónicas, situadas en diferentes regiones de su espectro.

8. **Estridente/mate.**- Afectan únicamente a las consonantes. Las consonantes que poseen el rasgo estridente se caracterizan por una distribución desigual, desordenada, de las áreas de sus frecuencias. El rasgo mate aporta espectrogramas con áreas de frecuencias que presentan estriaciones bien verticales bien horizontales. Cuando son horizontales, recuerda, salvando su inarmonicidad a los fonemas vocálicos.

9. **Bloqueado o glotalizado/no bloqueado o no glotalizado.**- Los sonidos que poseen el rasgo bloqueado se caracterizan por una proporción elevada de descarga de energía en un intervalo reducido de tiempo. El rasgo no bloqueado se caracterizan por una proporción más baja en esta descarga, en un intervalo más largo.

2.1.4.2. Rasgos de tonalidad

1. **Grave/no grave**.- El rasgo grave se manifiesta por medio de una concentración de energía en las frecuencias bajas, mientras que el rasgo no grave presenta concentraciones de energía en las zonas altas y en las medias.

2. **Agudo/no agudo**.- El rasgo de agudeza se caracteriza por una concentración de energía en la zona de frecuencias altas, y el rasgo opuesto presenta la concentración de energía en las frecuencias medias.

3. **Bemolizado/no bemolizado o normal**.- Acústicamente, el rasgo de bemolización se hace patente en los espectrogramas por el descenso en la línea de frecuencias de alguno o de todos los formantes. El rasgo normal se manifiesta por lo contrario.

4. **Sostenido/no sostenido**.- El rasgo sostenido viene determinado por una elevación del segundo formante, o por un refuerzo de alguno de los componentes de alta frecuencia. El rasgo no sostenido no acusa estos cambios.

2.1.5. Percepción de los sonidos

Veremos en este apartado varios aspectos relativos a la habilidad de los oyentes humanos para percibir y discriminar los sonidos.

2.1.5.1. Bandas críticas

El concepto de bandas críticas está asociado al fenómeno del enmascaramiento. Cuando dos tonos adyacentes se escuchan simultáneamente, el umbral de amplitud para percibir el tono más débil aumenta, es decir, el tono se enmascara por el otro, que es perceptualmente dominante. Se comprueba experimentalmente que el umbral aumenta solamente si la diferencia en frecuencia entre los dos tonos está por debajo de un valor crítico. Más allá de este valor crítico la percepción del primer tono no se ve afectado por el segundo.

Estos experimentos conducen al concepto de bandas críticas. Un sonido cuya frecuencia esté dentro de una banda crítica puede influir en la percepción de otro sonido en la misma banda, pero no de fuera de ésta. Experimentalmente se comprueba que el ancho de una banda crítica aumenta según sea mayor la frecuencia central de la ésta. Así para una frecuencia central del 200 Hz el ancho de banda es de unos 100 Hz y para 5kHz es de 1 kHz.

Desde un punto de vista psicológico, los filtros de banda crítica pueden considerarse como filtros de banda de paso cuya frecuencia se corresponde más o menos con las curvas de ajuste de las neuronas del sistema auditivo. Dos sonidos en la misma banda excitarán a las mismas neuronas y cada una interferirá en la percepción del otro.

Una medida perceptual, conocida como *bark*, asocia la frecuencia absoluta de un sonido y la resolución en frecuencia del oído en términos de banda crítica. Así un bark cubre el rango de frecuencias de una banda crítica. Su expresión analítica puede encontrarse en [ZWI80].

De manera similar, encontramos la escala mel. Esta escala define una relación entre el tono subjetivo y la frecuencia, donde la escala se ajusta para que 1000 mels

correspondan a 1kHz. La correspondencia es lineal por debajo de este valor y logarítmica por encima.

2.1.5.2. Altura

La altura de un sonido depende la vez de la intensidad y la frecuencia. De hecho la altura es el correlato subjetivo de la intensidad, de la misma manera que el tono es correlato subjetivo de la frecuencia.

La altura es una función no lineal de la intensidad. Las curvas de misma altura se establecen comparando sonidos puros de diferentes amplitudes y frecuencias.

2.1.5.3. Saturación, adaptación, enmascaramiento, supresión e inhibición lateral

Revisaremos ahora de forma breve algunos de los mecanismos no lineales asociados con la audición y que han sido empleados en el diseño de modelos acústicos.

La respuesta de las fibras nerviosas del sistema auditivo depende de la duración e intensidad del estímulo. Para una duración dada, la respuesta se incrementa con la intensidad o el nivel de presión sonora hasta un cierto umbral de saturación. Ese es el fenómeno de saturación.

Para una nivel de presión sonora, la respuesta decrece regularmente cuando la duración del estímulo crece y se acerca asintóticamente a un valor estable. Este es el fenómeno de la adaptación a corto plazo, que desempeña un papel importante en la percepción de cambios rápidos en frecuencia y amplitud.

La importancia del enmascaramiento se ha mostrado en la presentación de las bandas críticas. Otro aspecto relacionado con el enmascaramiento es el fenómeno de la supresión de dos tonos [JAV83]. Experimentalmente se observa que la respuesta de los filtros neuronales auditivos a un tono puro cuya frecuencia es igual a la frecuencia característica de esos filtros puede disminuir por la presencia de otro tono puro, incluso cuando el tono por separado no produzca excitación en los citado filtros. De acuerdo con este fenómeno, los componentes de alta frecuencia de un sonido complejo son los

más influyentes en la respuesta de las fibras nerviosas.

Otro fenómeno importante es el enmascaramiento hacia adelante de las fibras nerviosas auditivas. Consiste en la disminución en la respuesta a un determinado sonido debido al sonido precedente por lo general más intenso. El enmascaramiento hacia atrás también ha sido propuesto [ELL62].

Existen otras relaciones temporales entre los estímulos y el disparo de las fibras nerviosas. De manera precisa, la información acerca del estímulo se incorpora en el esquema temporal de disparos de los nervios auditivos. En respuesta a un tono puro los disparos de las fibras tienden a sincronizarse con el estímulo. Las fibras nerviosas no se disparan necesariamente en cada ciclo del estímulo, pero cuando lo hacen se produce en una única fase del estímulo. Sin embargo este fenómeno no se produce en todo el rango de frecuencias audible, estado situado el límite superior alrededor de los 4-5kHz.

Por ultimo, otro fenómeno que puede desempeñar un papel relevante en la percepción del habla es la inhibición lateral [HOU71]. Se puede describir como la supresión de la actividad de las fibras nerviosas de la membrana basilar causadas por la actividad de las fibras adyacentes. La inhibición lateral puede contribuir a la alta selectividad con respecto a la frecuencia del sistema auditivo.

2.2. CLASIFICACIÓN DE LOS SONIDOS BASADA EN PROPIEDADES ESPECTRALES

2.2.1. Vocales

El grupo de sonidos vocálicos es el que más interés ha despertado en las investigaciones acústicas por su indudable complejidad teórica y práctica. La dicotomía vocal/consonante es una noción básica que está presente en los estudios del lenguaje desde la antigüedad.

A grandes rasgos, el reconocimiento de una consonante a través de su percepción depende esencialmente de la presencia de un cambio de frecuencias en sus elementos acústicos constitutivos, mientras que el de una vocal depende de la estabilidad en la frecuencia.

Las vocales se caracterizan por presentar en su espectro una serie de formantes claramente definidos. Los pulsos glotales producidos en las cuerdas vocales apenas encuentran obstáculos en su camino de salida; pero el conjunto de cavidades supraglóticas a través del cambio en su forma y/o dimensión, hace que se resalten o se atenúen los diferentes armónicos del tono fundamental (frecuencia de vibración de las cuerdas vocales). El tracto vocal se comporta, por tanto, como una caja de resonancia. Cada conjunto de armónicos reforzados como consecuencia de este fenómeno es un formante, que en definitiva conforma el conjunto de características del timbre de una vocal [LAD96], [JOH97].

De todos los formantes, los dos primeros y en ocasiones el tercero son indispensables para la percepción de cada vocal. El resto contribuyen sobre todo a la caracterización del timbre individual de las mismas.

En cuanto a la percepción, intervalos iguales de frecuencia no guardan una correlación exacta con intervalos iguales de tonía; la percepción de diferencias cualitativas no se refiere simplemente a los intervalos de la frecuencia de los formantes.

La escala óptima de percepción es aproximadamente lineal en las bajas frecuencias (por debajo de 1000 Hz) y logarítmica en el resto.

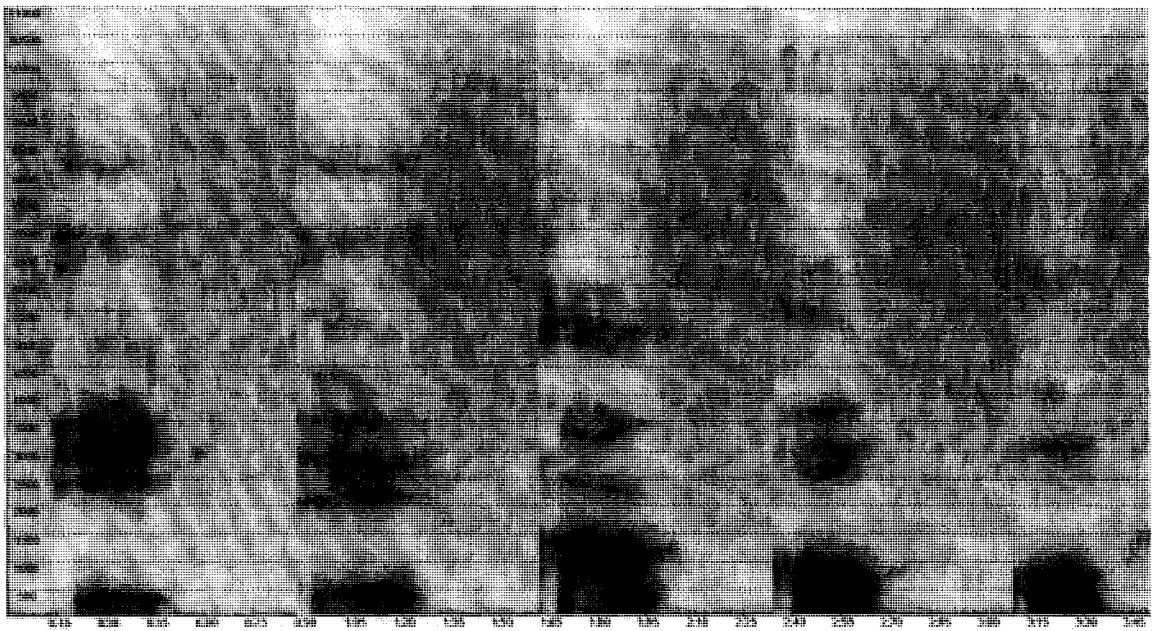

Figura 2.4. Espectrograma de los sonidos vocálicos [a], [e], [i], [o], [u] realizados por un hablante masculino.

Podemos establecer algunas relaciones entre las frecuencias formánticas y las configuraciones de la cavidad bucal:

a).- Cuanta más alta es la frecuencia del primer formante F_1 mayor es la abertura total de la cavidad oral y viceversa.

b).- Cuanto más posterior sea la posición de la lengua o cuanto mayor sea el redondeamiento y la proyección labial, más baja será la frecuencia del F_2, y a la inversa.

c).- Existe una relación directa entre la elevación del F_3 y el descenso del velo del paladar, como en la nasalización, y entre el descenso del F_3 y la elevación de la punta de la lengua hacia una posición retrofleja.

Existen dos hechos importantes respecto a las posiciones de los formantes vocálicos:

- La identificación lingüística de las vocales no depende enteramente de la frecuencia absoluta de los formantes del sujeto hablante, estructura que puede

variar ligeramente de una persona a otra, especialmente entre hombres, mujeres y niños.

Vocal /í/			**F_1**	**F_2**
1.	[bíβo]	vivo	202 Hz	2308 Hz
2.	[akí]	aquí	202 Hz	2632 Hz
3.	[amaríʎos]	amarillos	202 Hz	2592 Hz
4.	[día]	día	202 Hz	2511 Hz
5.	[θeríʎas]	cerillas	243 Hz	2551 Hz
Vocal /é/			**F_1**	**F_2**
6.	[béβe]	bebe	324 Hz	2146 Hz
7.	[θerβeθa]	cerveza	283 Hz	2025 Hz
8.	[péro]	perro	405 Hz	1822 Hz
9.	[néɣras]	negras	283 Hz	2349 Hz
10.	[tréθe]	trece	283 Hz	2106 Hz
Vocal /á/			**F_1**	**F_2**
11.	[báβa]	baba	729 Hz	1174 Hz
12.	[káða]	cada	648 Hz	1417 Hz
13.	[pa̱ra]	parra	607 Hz	1012 Hz
14.	[táca]	tacha	769 Hz	1417 Hz
15.	[baʝa]	vaya	648 Hz	1093 Hz
Vocal /ó/			**F_1**	**F_2**
16.	[bóβo]	bobo	405 Hz	850 Hz
17.	[jo]	yo	324 Hz	931 Hz
18.	[tóðo]	todo	283 Hz	972 Hz
19.	[óco]	ocho	283 Hz	891 Hz
20.	[dóθe]	doce	364 Hz	1012 Hz
Vocal /ú/			**F_1**	**F_2**
21.	[be̯intiúno]	veintiuno	202 Hz	567 Hz
22.	[birtú]	virtud	202 Hz	729 Hz
23.	[sepultúra]	sepultura	202 Hz	850 Hz
24.	[púpas]	pupas	243 Hz	688 Hz
25.	[r̄úso]	ruso	243 Hz	769 Hz

Tabla 2.4. Ejemplos de realizaciones de fonemas vocálicos tónicos en sílaba libre de un informante femenino [QUI88].

- Además, dentro de un mismo hablante, los valores de frecuencia para los formantes de un fonema vocálico son diferentes debido a fenómenos de coarticulación. Como ejemplo en la Tabla 2.4 se ofrece una serie de realizaciones, pertenecientes todas ellas a fonemas vocálicos tónicos en sílaba libre de un informante femenino.

Las modificaciones frecuenciales de los tres primeros formantes pueden dar cuenta también de hechos articulatorios generales de los sonidos que preceden a sonidos vocálicos:

a).- La labialización se caracteriza por el descenso en su gama de frecuencias del F_2, sobre todo, y del F_3.

b).- La palatalización se caracteriza por un considerable aumento de la frecuencia del F_2, un pequeño aumento del F_3 y un ligero descenso del F_1.

c).- La velarización se manifiesta por un amplio descenso de las frecuencias del F_2 y prácticamente insignificante del F_3, permaneciendo casi inalterado el F_1.

La clasificación acústica de las vocales puede establecerse mediante los rasgos: vocálico, continuo, sonoro y no consonántico. Los otros rasgos que caracterizan las vocales entre sí, son: denso-difuso, nasal-oral, grave-agudo, bemolizado-no bemolizado.

El rasgo vocálico, acústicamente se manifiesta en que:

a).- Presenta una estructura de formantes definida, resultando la mayor concentración de energía en la región entre los 300 y los 800 Hz.

b).- Los formantes altos de las vocales están menos atenuados que los otros sonidos que posean una estructura formántica.

c).- Los sonidos vocálicos tienen, en general, mayor intensidad que los consonánticos; solamente superados por los sonidos [s], [c] (momento fricativo) y [r], aunque este último no en todos los casos.

El rasgo denso o compacto se manifiesta, en la práctica por la situación frecuencial del F_1 respecto de F_2. Cuanto más alto y más cerca esté F_1 de F_2, más compacta será la vocal. En español es compacta la /a/, siendo no compactos y difusos los fonemas /i/, /u/ y no compactos y no difusos /e/, /o/.

El rasgo nasal por su parte se caracteriza en las vocales por una reducción de la intensidad del F_1. En español, este rasgo no es pertinente, apareciendo cuando la vocal está situada entre dos consonantes nasales, o bien, en posición inicial absoluta, seguida

de consonante nasal.

El rasgo grave-agudo se manifiesta en las vocales como consecuencia de la posición del F_2. La altura en frecuencia del segundo formante es directamente proporcional al grado de agudeza, e inversamente proporcional al de gravedad. Así, en español, son graves: /o/, /u/, son no graves y agudas /i/, /e/ y es no grave y no aguda /a/.

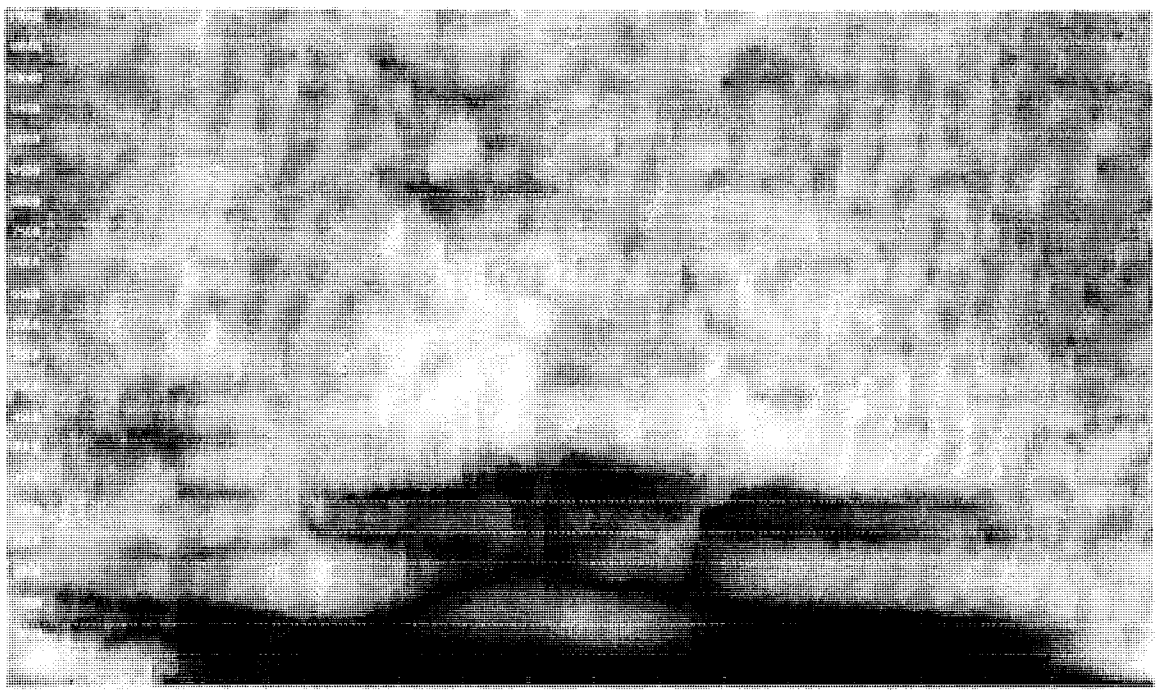

Figura 2.5. Espectrograma de [aωélo] *abuelo* realizada por un hablante masculino. Puede observarse la semiconsonante posterior [ω] entre los instantes de tiempo 0,10-0,24s.

2.2.2. Oclusivas orales

Desde el punto de vista acústico, las consonantes oclusivas o explosivas orales se diferencian del resto por los siguientes motivos:

- La interrupción total en la emisión de sonido, que se produce durante la tensión de la consonante.
- La explosión que sigue a esta interrupción. Esta explosión se manifiesta en forma de sonido turbulento, breve e intenso.
- La rapidez de las transiciones de los formantes de las vocales precedentes o siguientes.

Los espectrogramas de las oclusivas sordas se caracterizan por la ausencia total de zonas de frecuencia; en las sonoras, esta ausencia también es patente, pero una barra de sonoridad en la parte inferior de su espectro, producto de la vibración de las cuerdas vocales, las diferencia de las anteriores.

En sí, el espectro de las explosivas tanto sordas como sonoras, no proporciona ningún dato que las caracterice y permita separarlas del resto de su grupo. Las dos fuentes de información que permiten clasificar entre sí a las oclusivas son:

1. El lugar de la explosión, es decir, la frecuencia donde se concentra la energía en la barra de explosión que está situada al final del segmento explosivo e inmediatamente antes de la vocal siguiente.
2. Las transiciones entre la oclusiva y la vocal siguiente.

El primer factor por sí solo no suministra ningún dato que pueda caracterizar una consonante explosiva, puesto que dos frecuencias de explosión pueden asociarse a un mismo sonido y varios sonidos pueden presentar la misma frecuencia de explosión. En cualquier caso este resultado es dependiente de la vocal que acompaña al sonido oclusivo.

El segundo factor se refleja en los espectrogramas como movimientos rápidos y

continuos en los formantes. Esto es fruto del movimiento articulatorio hacia la abertura combinado con un desplazamiento complejo de los órganos. La transición del primer formante T_1 ayuda en la percepción de la sonoridad: cuanto más débil sea, tanto más sorda se percibe, y viceversa. La transición del segundo formante T_2 es el mejor índice para la localización del lugar de articulación, aunque este resultado sea dependiente de cada vocal. La contribución de T_3 en la percepción de los sonidos disminuye desde los dentales a los labiales, siendo aún menor en los velares.

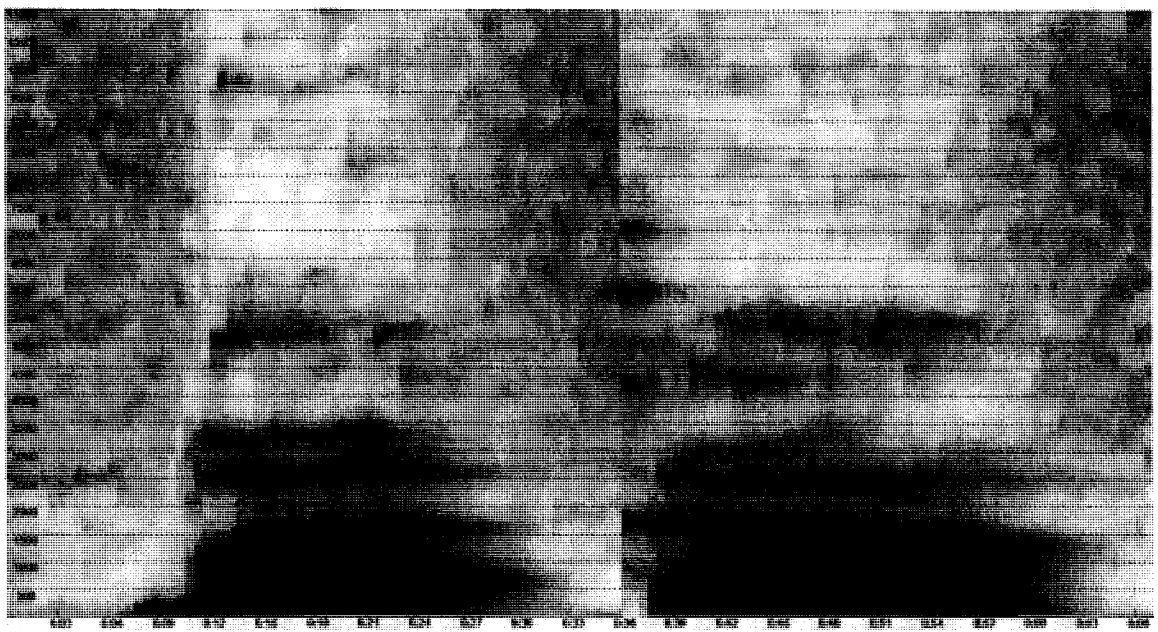

Figura 2.6. Espectrograma de [báta] *bata* realizada por un hablante masculino. Encontramos dos sonidos oclusivos bilabiales: uno sonoro [b] (instantes 0,07-0,12s.) y otro sordo [t] (instantes 0,30-0,36s.) en posición intervocálica. Para el primero se observa la barra de sonoridad situada a una frecuencia de unos 200 Hz.

Las transiciones de sus formantes segundo y tercero para el caso del castellano, vienen reflejadas en la Tabla 2.5:

[p]	-	-	-	-	-	-	-	-	-	-
[t]	-	-	-	-	+	-	+	-	+	-
[k]	-	-	+	+	+	-	+	-	=	-
[b]	-	-	-	-	-	-	-	-	-	-
[d]	-	-	-	-	+	-	+	-	+	-
[g]	=	+	+	+	+	=	=	-	=	-

Tabla 2.5. Transiciones de los formantes F_2 y F_3 para las oclusivas orales del español [QUI88].

2.2.3. Oclusivas nasales

La mayor diferencia entre oclusivas nasales y las orales es la existencia en las primeras de ciertos formantes durante su momento de tensión. Estas resonancias sirven, aunque en un grado pequeño, para identificar el lugar de articulación. A pesar de todo son las transiciones con las vocales contiguas, los índices que permiten identificar los lugares de articulación. En este sentido comparten con las oclusivas orales la forma y dirección de las transiciones del segundo y tercer formante de las vocales continuas.

Desde el punto de vista fonológico, el español conoce tres fonemas nasales /m/, /n/, /ɲ/, que funcionan como tales en posición silábica prenuclear. En posición silábica postnuclear quedan neutralizados, dando como resultado el archifonema nasal /N/. Para este último caso su lugar de articulación no es pertinente y lo importante es que se realice una oclusión bucal y que quede una resonancia nasal.

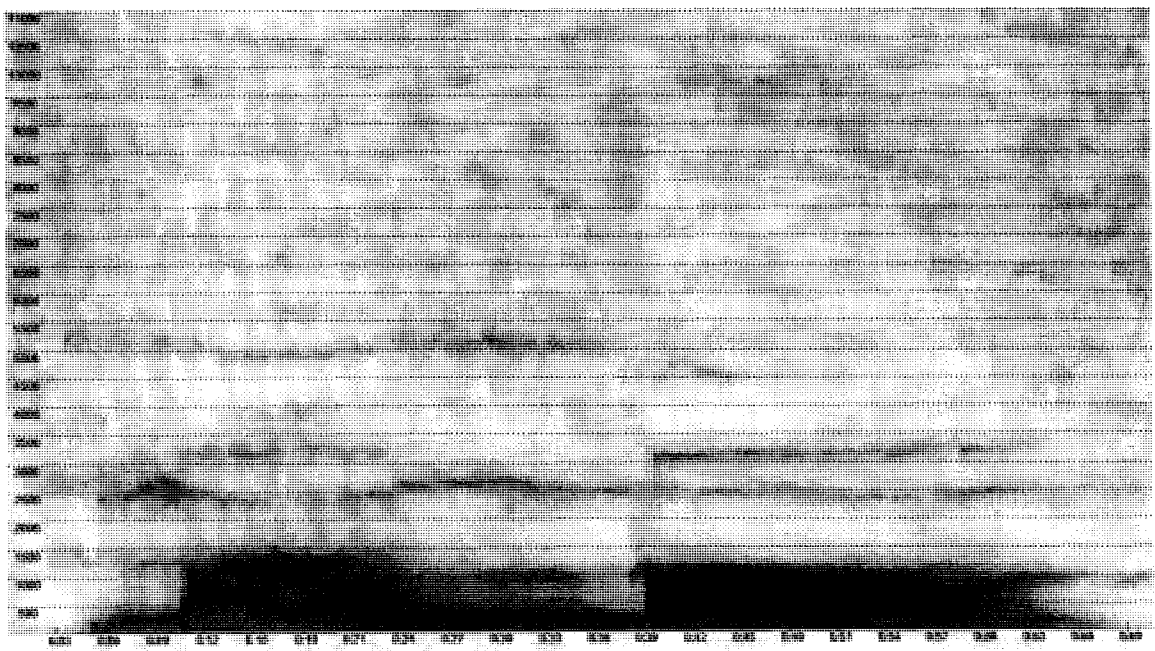

Figura 2.7. Espectrograma de [mámbo] *mambo* realizada por un hablante masculino. Aquí encontramos dos realizaciones del sonido nasal bilabial sonoro [m] en los instantes de tiempo 0,05-0,10s. y 0,24-0,38s. respectivamente.

Las transiciones en este grupo de sonidos son:

- Las transiciones del F_1 son siempre negativas en todos los casos.

- Las transiciones del F_2, son negativas para /m/ en todos los casos y para /n/ con /i/, /e/, /a/. Son positivas para /n/ con /o/, /u/ y para /ɲ/ en todos los casos excepto con /i/ en que no existe transición.

- Para el F_3 encontramos transiciones negativas con /i/, /e/ y negativas con /o/, /u/ excepto para /m/ que no presenta transición. El caso de /a/ no presenta transiciones. En todas las situaciones la transición T_3 para /ɲ/ es muy pequeña fluctuando entre positiva y negativa.

Acústicamente, los fonemas oclusivos nasales del español son nasales, sonoros, continuos y difusos (F_1 aparece en zonas muy bajas del espectro y F_2, cuando lo hace en zonas superiores a los 1000 Hz). Además /m/ es grave, /ɲ/ es agudo no grave y /n/ es no grave y no agudo.

2.2.4. Fricativas

Las consonantes que se engloban dentro de este grupo reciben el nombre de fricativas, por ser la fricción que produce el aire al pasar a través de la estrechez formada entre dos órganos articulatorios el rastro más notoriamente audible en ellas. El elemento más perceptible de las consonantes fricativas se encuentra en su tensión, tanto acústica como articulatoriamente.

En la caracterización acústica de las fricativas destaca el ruido de fricción que se refleja en las partes altas del espectro. Además, influyen en los formantes de las vocales contiguas introduciendo ciertas transiciones. El ruido de fricción no basta para la identificación de todas las fricativas sino que además se requiere conocer las transiciones de segundos y terceros formantes.

Las fricativas, en español presentan una pequeña irregularidad: en el plano fonológico existen cinco fonemas: /f/, /θ/, /s/, /ʃ/, /x/; pero en el plano acústico hay además tres más sonidos más: [β], [ð], [γ] que son alófonos, en distribución complementaria de /b/, /d/, /g/. Aparte de esto, las consonantes fricativas, presentan en español dos grupos bien delimitados: a) el de aquellas consonantes que poseen predominio de resonancias en las zonas de bajas frecuencias, y b) el de las fricativas cuyas resonancias se encuentran en las zonas altas de frecuencias u ocupan todo el espectro.

2.2.4.1. Fricativas de resonancias bajas

Se encuadran en este grupo a [β], [ð], [γ], [ʃ]. Se caracterizan porque las zonas de resonancia se encuentran situadas, principalmente, en la mitad inferior de su espectro.

Los alófonos fricativos [β], [ð], [γ] se distinguen principalmente de los oclusivos correspondientes por la presencia en el espectro de los primeros de las zonas de frecuencia más o menos amplias e intensas, que se aproximan en su configuración a los formantes vocálicos. La aparición de estas zonas de resonancia depende del grado de constricción de los órganos articulatorios. El movimiento de los formantes es aproximadamente el mismo que el que poseen las explosivas, con algunas pequeñas

diferencias en cuanto a la velocidad de la transición T_1, o en las transiciones T_2 y T_3 en algunos casos.

El sonido [ĵ] se caracteriza por presentar un F_1 mucho más bajo que el de las vocales contiguas y un F_2 más débil que el de éstas. Las transiciones son lentas siendo la transición del segundo formante muy positiva.

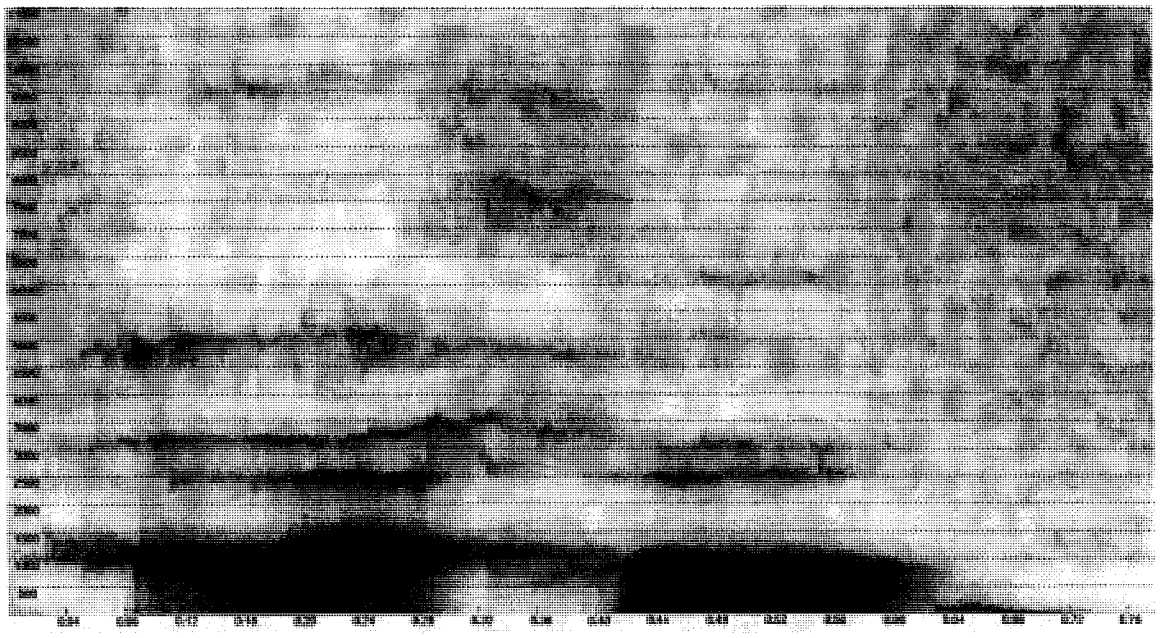

Figura 2.8. Espectrograma de [aβáχo] *abajo* realizada por un hablante masculino. Entre los instantes de tiempo 0,10-0,20s. se aprecia el sonido fricativo bilabial sonoro [β].

2.2.4.2. Fricativas de resonancias altas

Encontramos en este grupo los fonemas /f/, /θ/, /s/ /x/. De los cuatro el que presenta mayor intensidad es el /s/ seguido de /x/ y por último /f/, /θ/.

El sonido [f] por su carácter labial, se caracterizar por poseer unas transiciones T_2 y T_3 negativas en contornos de /i/, /e/, /a/, y horizontal con los demás sonidos vocálicos. Su frecuencia comienza a unos 2100 Hz.

El sonido [θ] tiene una frecuencia de aparición que varía desde los 2400 Hz para [θu] hasta los 6000 Hz de [θa], debido al aumento de volumen de la cavidad anterior de resonancia, motivado por la labialización. Las transiciones de los formantes F_2 y F_3 son negativas con [a, e, y]. Con [o, u] T_2 es positiva y T_3 es negativa.

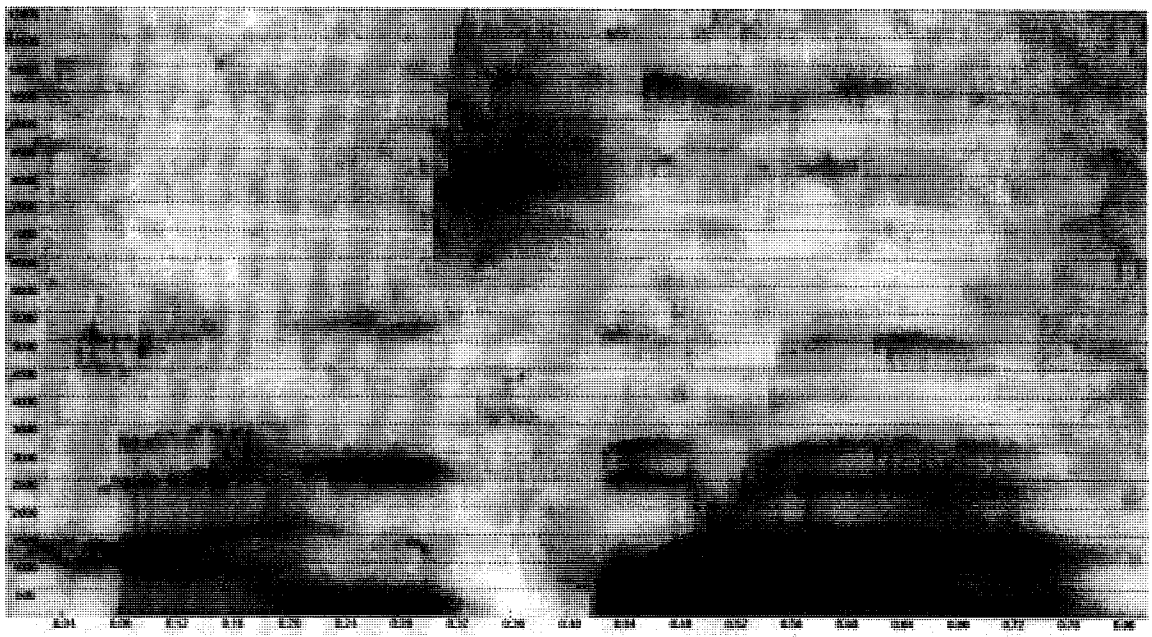

Figura 2.9. Espectrograma de [ánfora] *anfora* realizada por un hablante masculino. Entre los instantes de tiempo 0,32-0,44s. se aprecia claramente el sonido fricativo labiodental sordo [f], estando situado en frecuencia por encima de los 6500 Hz.

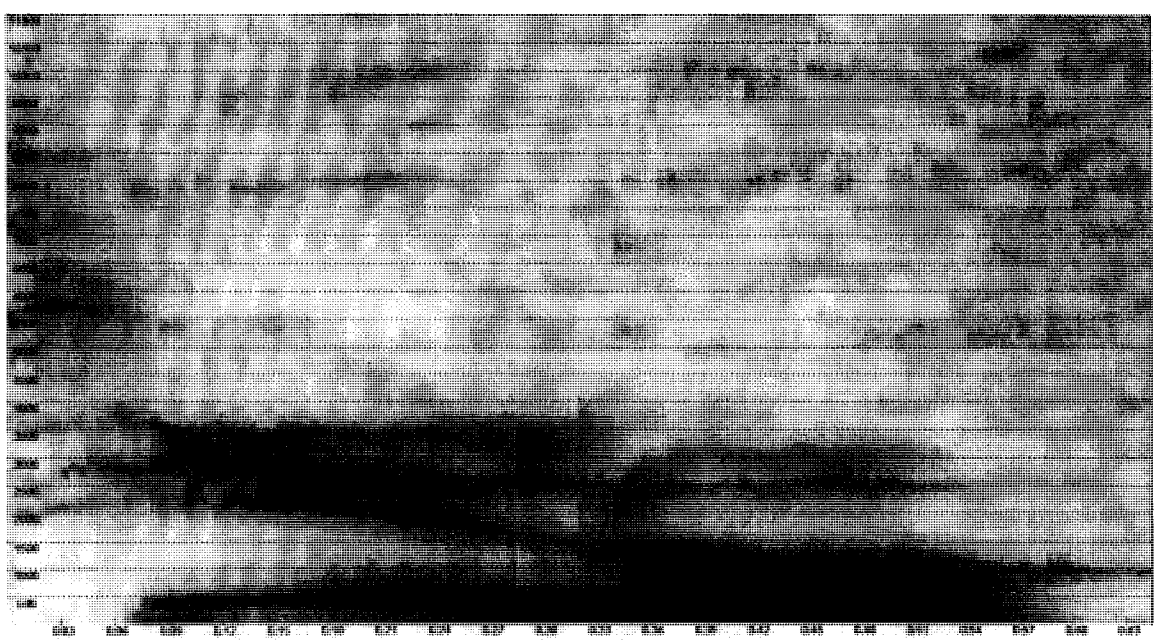

Figura 2.10. Espectrograma de [θjélo] *cielo* realizada por un hablante masculino. Puede apreciarse entre los instantes de tiempo 0,0-0,07s. el sonido fricativo linguointerdental sordo [θ].

El fonema /s/, en español, es uno de los que tiene mayor número de realizaciones. La frecuencia a la que se inicia la banda de fricación va desde los 2500 Hz hasta los 5700 Hz dependiendo de la articulación (cuanto más anterior, mayor frecuencia de

comienzo). Aún así la frecuencia de comienzo aparece más alta con [i, e, a] y más baja con [o, u]. Cuanto más posterior sea la realización, tanto más estridente es la realización. Sus transiciones varían mucho según la articulación.

Por último el sonido [x] presenta el comienzo de su banda de fricación en una altura bastante variable. Sus transiciones tienen los siguientes movimientos: T_2 y T_3 son negativas con [i, e] y positivas con [a]; con [o, u] es positiva T_2 y negativa T_3.

La clasificación acústica de las consonantes fricativas se realiza del siguiente modo:

- Son todas orales y continuas.
- Son sonoras: [β, ð, γ, ʝ] y sordas el resto.
- Son compactas: [ʝ, x, γ, ç, h] y difusas: [f, θ, s, ð, β].
- Son graves: [f, β, x, h] y agudas: [θ, s, ð, ʝ, ç].
- Es estridente: [s] y mates el resto.

2.2.5. Africadas

Las consonantes africadas se caracterizan porque en su emisión intervienen dos momentos: uno interrupto, similar al de las explosivas, seguido de otro constrictivo. Estos dos momentos se realizan en el mismo lugar articulatorio y, además, durante el momento de su tensión. Por ello, los espectrogramas participan de las características de los dos modos antes descritos.

El momento de constricción de la africada es mayor que el que se consigue en una explosiva aspirada (alrededor de 30 ms.) y normalmente menor que el de una fricativa (aproximadamente 110 ms.).

El castellano presenta un único fonema africado /c/, pero fonéticamente posee dos africadas: una, el alófono de /c/, y otra la palatal sonora, alófono del fonema fricativo linguopalatal central /ʃ/ que se transcribe por /ɟ/.

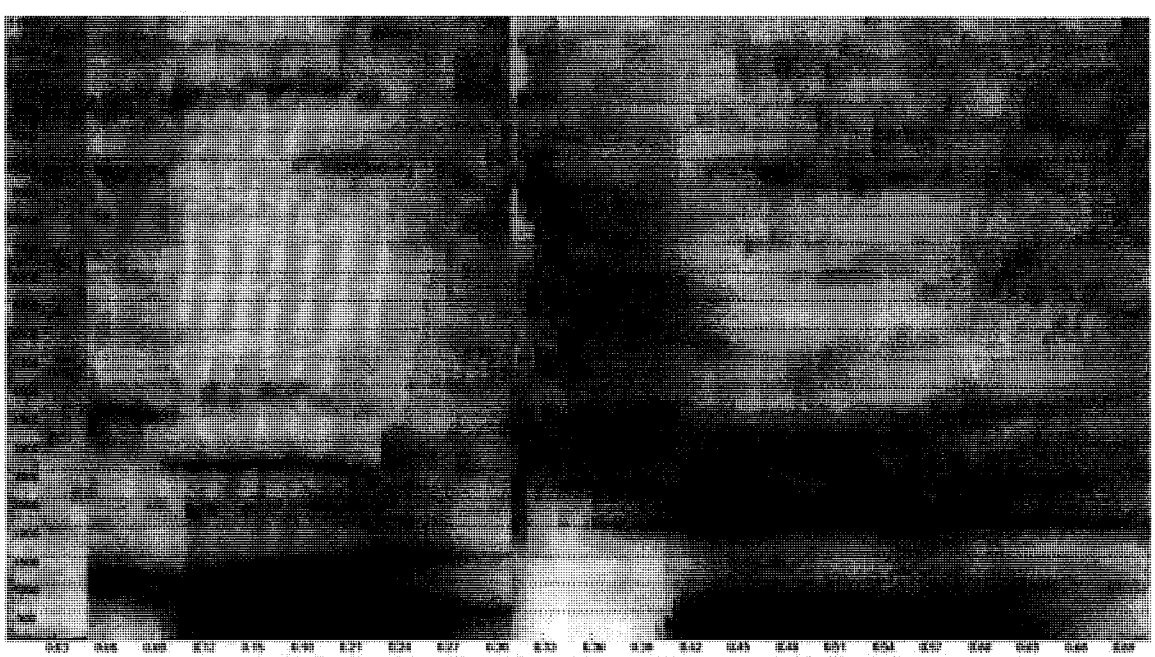

Figura 2.11. Espectrograma de [kóce] *coche* realizada por un hablante masculino. Puede observarse el sonido africado linguopalatal sordo [c] (instantes de tiempo 0,30-0,42s.). Dentro de este sonido destaca la presencia de los dos momentos descritos anteriormente (oclusivo y fricativo).

Tanto para un caso como para el otro, el ruido de fricación suele comenzar a la altura del comienzo de la transición del F_2. Las transiciones se caracterizan por tener un

T_1 negativo. T_2 es negativo con vocales agudas [i, e], positivo con [a] y de un grado positivo muy acusado con [o, u]. Por su parte T_3 es claramente negativo con [o, u]; menos acusadamente negativo con [a], y aparece sin transición o con transición positiva con [i, e].

Acústicamente nuestro fonema africado es compacto, agudo, interrupto y estridente.

2.2.5.1. Africadas sordas

En las africadas sordas, normalmente la oclusión es mayor que la fricación; pero ocurre a veces, en ciertas áreas dialectales, que la oclusión es menor: en estos casos se puede vislumbrar una tendencia hacia la pérdida del momento oclusivo, y a convertir la consonante explosiva en fricativa.

La frecuencia a la que aparece la fricación nos indicará el lugar de articulación: su altura es directamente proporcional a la anterioridad articulatoria.

2.2.5.2. Africadas sonoras

Se diferencian de las anteriores fundamentalmente por su longitud que es menor en la mayoría de los casos y por la aparición de la banda de fricación en frecuencias menores.

2.2.6. Líquidas

El grupo de consonantes líquidas, del que se han realizado muy pocos estudios acerca de la naturaleza acústica, engloba a las consonantes laterales y vibrantes. Estos sonidos presentan unas características que les infieren una fisionomía intermedia entre los sonidos vocálicos y los consonánticos. La estructura formántica de las líquidas es muy similar a la de las vocales, pero difiere de éstas en:

a).- La frecuencia fundamental es menor.

b).- Su intensidad global es menor también.

2.2.6.1. Líquidas laterales

Se caracterizan por su continuidad, lo que da origen a que en su espectro aparezcan ciertos formantes análogos a los vocálicos. En español, existen dos fonemas laterales: /l/, /λ/.

Además presentan unas características algo diferentes dependiendo de si se encuentran en sílaba tónica o átona:

a).- Presente en sílaba tónica aparece mayor duración que si se encuentra en sílaba átona.

b).- Los valores de F_1 son menores en posición tónica y los de F_2 y F_3 menores en posición átona.

En el sonido /l/ del español, el F_1 aparece a una frecuencia más baja que el correspondiente de la vocal con la que forma sílaba y se ve poco influido por ella. El valor medio puede establecerse en 320 Hz.

El F_2, aparece aproximadamente a la altura de la /a/ (1575 Hz) y presenta un cierto movimiento hacia un punto situado en el comienzo de la transición de su vocal silábica. Sus transiciones son negativas con /i/, /e/ y positivas con /o/, /u/; con /a/ pueden darse los tres casos aunque generalmente la transición es ligeramente positiva. La intensidad de /l/ es menor que la de su vocal silábica.

El sonido /λ/ presenta valores para el F_1 (290 Hz de media), menores que los correspondientes al primer formantes de /l/ y no hay diferencia entre posiciones tónicas y átonas.

El F_2 (2050 Hz de media) y el F_3 (2650 Hz de media) mantienen una frecuencia más alta en sílaba tónica y en cualquier caso, sus frecuencias son mayores que las que presenta el sonido /l/ para los respectivos formantes.

Las transiciones del primer formante son siempre negativas. Las transiciones del F_2 son siempre positivas con /a/, /o/, /u/, predominando con /i/, /e/ las de este signo.

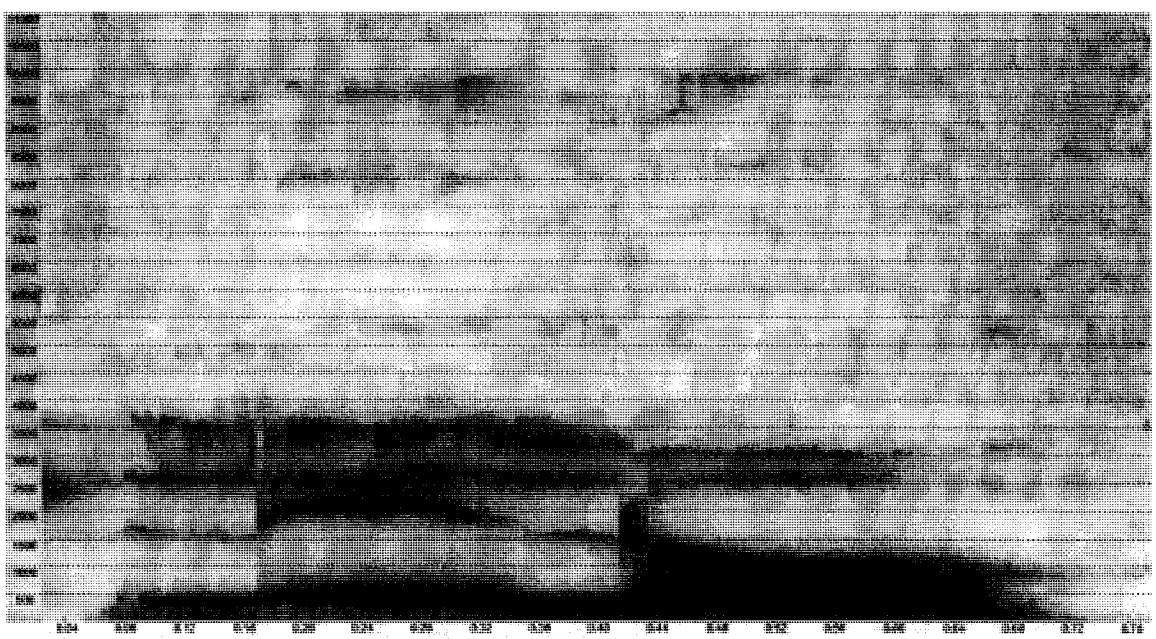

Figura 2.12. Espectrograma de [lélo] *lelo* realizada por un hablante masculino. En la representación se observa claramente las zonas correspondientes a los dos sonidos laterales linguoalveolares sonoros [l] (instantes 0,08-0,16s. y 0,32-0,42s. respectivamente). Estas zonas presentan una estructura de formantes claramente diferenciada de las regiones vocálicas vecinas.

2.2.6.2. Líquidas vibrantes

Las consonantes vibrantes se caracterizan por su cualidad de interruptas. En español encontramos dos fonemas vibrantes: /r/ y /r̄/.

La /r/ presenta una interrupción breve producto de una rápida oclusión articulatoria del ápice de la lengua contra los alveolos. Las transiciones de los formantes adyacentes a esta consonante son las siguientes: T_1 es negativa; T_2 es negativa con /i/,

/e/, positiva con /o/, /u/; con /a/ la transición es pequeña a veces positiva a veces negativa; T_3 con /a/ es muy pequeña tanto si negativa como positiva, y negativa con las demás vocales.

La vibrante múltiple /r̄/, presenta un número medio de interrupciones de tres con dos momentos vocálicos, que corresponden a los momentos de abertura entre el ápice lingual y los alveolos. Sus transiciones son las siguientes: T_1 es siempre negativa. T_2 es negativa con /i/, /e/, /a/ y positiva con /o/, /u/. T_3 es pequeña con /a/ tanto positiva como negativa y negativa con las demás vocales.

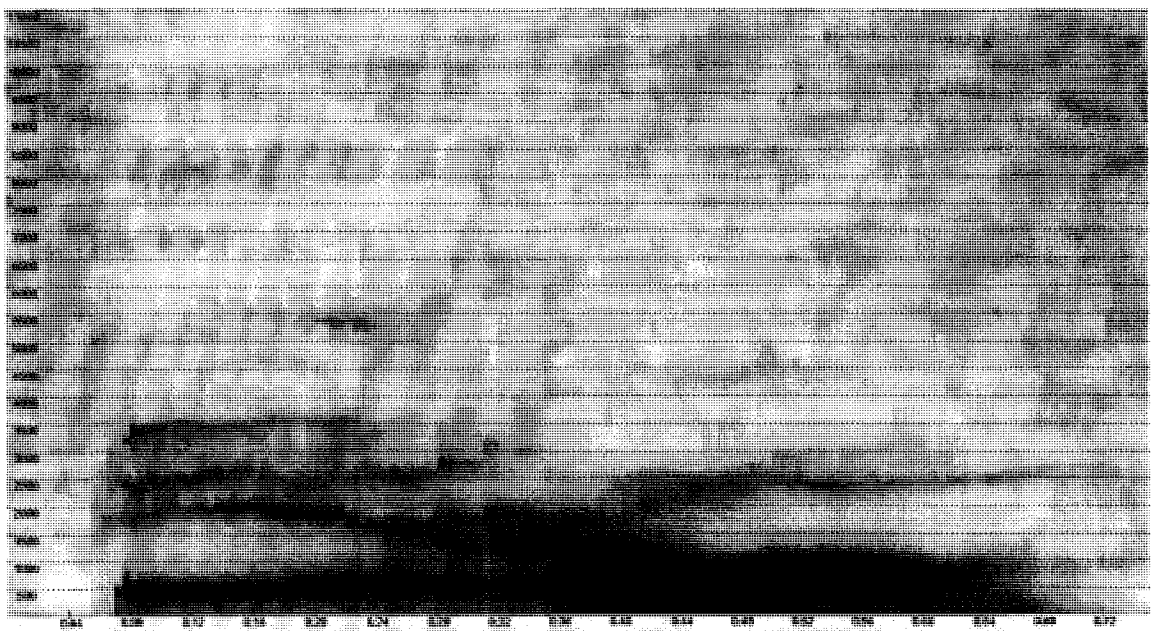

Figura 2.13. Espectrograma de [pér̄o] perro realizada por un hablante masculino. El sonido vibrante múltiple linguoalveolar sonoro [r̄] (instantes 0.23-0,40s.) presenta tres interrupciones.

3. INTRODUCCIÓN AL RECONOCIMIENTO AUTOMÁTICO DEL HABLA

3.1. INTRODUCCIÓN

El reconocimiento automático del habla [COL97] es el procedimiento por el cual se convierte una señal acústica, capturada por un micrófono en conjunto de símbolos de un diccionario dado, asociados con elementos semánticos de tipo palabra, generalmente. Las palabras reconocidas pueden constituir el resultado final, como en aplicaciones de control, entrada de datos y preparación de documentos. También puede servir como entrada a otros módulos de procesamiento lingüístico con el fin de alcanzar la comprensión de la realización oral.

La forma de resolver el problema de reconocimiento suele adoptar una aproximación estadística basada en encaje de patrones al considerar que la señal de voz debe procesarse como si se tratase de un proceso estocástico.

Con esta aproximación se considera, que inicialmente se dispone de un modelo de generación de la fuente/canal de voz que produce una cierta secuencia de palabras W, es decir, la situación inversa a nuestro problema. Suponemos también, que este modelo se caracteriza por su falta de certeza y baja fiabilidad a la hora de convertir el conjunto de palabras anterior en la señal de voz S. De esta manera, modelamos la conversión de W en la señal observada S como si de un canal ruidoso se tratara. El reconocimiento se formula entonces como un problema de decodificación a posteriori en el cual debemos maximizar una determinada función objetivo [RAB96]. Una forma de simplificar el problema es no trabajar con S directamente sino con una representación paramétrica de S. La representación será una secuencia de vectores acústicos A. Si ahora empleamos la regla de Bayes para volver a formular el problema de decodificación, tenemos que:

$$\arg\max_{W \in \Gamma} P(W|A) = \arg\max_{W \in \Gamma} P(A|W) \cdot P(W) \tag{3.1}$$

donde Γ es el conjunto de todas las posibles secuencias de palabras, $P(A|W)$ es la probabilidad de la secuencia de vectores A, conocida la particular secuencia de palabras W y por último, $P(W)$ es la probabilidad a priori de generar la secuencia W. El primer término se conoce como el modelo acústico y al segundo $P(W)$ como al modelo de lengua.

Lo que pretendemos es modelar el sistema de producción de la voz como un canal ruidoso donde el ruido representa la variabilidad del hablante y del entorno. Teniendo en cuenta que no es posible en la práctica tener un conocimiento completo de las características de este canal, la aproximación estadística asume a menudo formas paramétricas particulares para $P_\theta(W|A)$ y $P_\omega(W)$, es decir, de acuerdo a modelos específicos. Todos los parámetros de los modelos estadísticos, como por ejemplo θ y ω, necesarios para la evaluación de la probabilidad acústica $P_\theta(W|A)$ y la probabilidad de lengua $P_\omega(W)$, se estiman a partir de largas colecciones, llamadas conjuntos de entrenamiento. Estas colecciones están formadas por un gran número de realizaciones de habla, producidas en general por el conjunto de hablantes representativo de los usuarios del sistema final y por grandes cantidades de texto escrito. Este proceso se conoce como entrenamiento de modelos o aprendizaje de modelos y constituye la base que garantiza el éxito del proceso de reconocimiento.

Los sistemas de reconocimiento han alcanzado un desarrollo notable en lo que se refiere a sistemas de laboratorio, aunque mucho menos en situaciones reales. La Tabla 3.1 resume los resultados de diversos reconocedores de voz, basados en evaluaciones de laboratorio, para las tres áreas típicas (palabras aisladas, conectadas y habla continua) y para diferentes tareas. La media de las prestaciones de los sistemas se expresa como la tasa de errores por palabra (en porcentaje) para un vocabulario y complejidad dadas.

Basándonos en el modelo específico de tareas, podemos encontrar cinco grandes categorías de aplicación de los sistemas de reconocimiento que cubren tanto las áreas de telecomunicación como de negocios:

a).- Telecomunicaciones. Proporciona información o acceso a datos y servicios sobre línea telefónica. Dos de las aplicaciones más extendidas incluyen el sistema de automatización para manejo de llamadas del servicio asistido por operador de AT&T o VRCP (*Voice Recognition Call Processing*) y el NTT ANSER para la realización de determinadas operaciones bancarias por teléfono.

b).- Oficina/escritorio. Para proporcionar capacidades de reconocimiento del habla en el manejo de ordenadores personales y sus aplicaciones, control de las

agendas telefónicas y funciones del teléfono, dictado y rellenado de formularios.

c).- Negocios y manufactura, permitiendo la incorporación de la voz en proceso de control de calidad en cadenas de montaje, empaquetado, ordenación y distribución.

d).- Área médica y legal. Generación de informes, documentos y formularios con presencia de abundante vocabulario específico.

e).- Otras aplicaciones como control domótico, ayuda a discapacitados, juegos, etc.

Tecnología	Tarea	Modo	Vocabulario	Tasa de error
Palabras aisladas	Palabras equiprobables	Dependiente del locutor	10 dígitos	0%
			39 caracteres alfanuméricos	4,5%
			1.109 palabras básicas del inglés	4,3%
		Independiente del locutor	10 dígitos	0,1%
			39 caracteres alfanuméricos	7,0%
			1.218 nombres de ciudades	4,7%
Palabras conectadas	Dígitos con longitud conocida	Dependiente del locutor	10 dígitos	0,1%
		Independiente del locutor	11 dígitos	0,2%
	Jerga de líneas aéreas (complejidad= 4)	Dependiente del locutor	129 términos	0,1%
Habla fluida	Resource Management (RM) (complejidad= 60)	Independiente del locutor	991 palabras	3,0%
	Airline Travel Information System (ATIS) (complejidad= 25)	Independiente del locutor	1.800 palabras	3,0%
	Wall Street Journal (WSJ) (complejidad= 145)	Independiente del locutor	20.000 palabras	12,0%

Tabla 3.1.Tasas de error por palabra para diversos sistemas de reconocimiento evaluados en laboratorio [RAB96].

3.1.1. Clasificación de los sistemas de reconocimiento de voz

El reconocimiento de la voz es un problema difícil de resolver, entre otras cosas porque existen muchas fuentes de variabilidad asociadas con la señal de entrada al sistema.

Las realizaciones acústicas de los diferentes fonemas, los alófonos, dependen mucho del contexto en el que aparecen. El problema es que pequeñas variaciones en estas realizaciones pueden dar lugar a fuertes cambios en el significado del mensaje:

a).- Los cambios en las condiciones del entorno de trabajo así como la calidad y posición de los micrófonos pueden ser origen también de variaciones acústicas.

b).- Las diferencias en el habla para un mismo locutor fruto de cambios en el estado de ánimo, ritmo de producción, o calidad de la voz en determinadas circunstancias como por ejemplo afecciones del aparato respiratorio.

c).- Las diferencias entre el habla de diversos locutores. En este caso las fuentes pueden ser el género, las diferencias sociolingüísticas, el dialectalismo e incluso el tamaño y forma del tracto vocal.

Debido a todo esto, un sistema de reconocimiento de voz puede caracterizarse por un conjunto de parámetros diverso. Entre los más importantes podemos citar:

Parámetros	Rango
Tipo de discurso	[Palabras aisladas, habla continua]
Dependencia del locutor	[Dependiente del locutor, independiente del locutor]
Tamaño del vocabulario	[Pequeño (< 20 palabras), Grande (> 20000 palabras)]
Estilo de discurso	[Lectura, habla espontanea]
Modelo de lengua	[Contexto explícito, sensible al contexto]
Confusión	[Pequeña (<10), grande (> 100)]
Relación señal ruido	[Alta (> 30 dB), baja (< 10 dB)]
Tipo de transductor	[Micrófono de gradiente, teléfono]

Tabla 3.2. Parámetros típicos empleados en la caracterización de un reconocedor de habla [COL97].

El rango expresa los límites de complejidad que tiene el sistema para un parámetro dado. La parte izquierda indica el caso más simple y la parte derecha la situación más compleja que se puede presentar.

Un sistema que reconozca palabras aisladas requiere que el hablante realice una pequeña pausa entre palabras, mientras que un sistema de reconocimiento continuo no. Algunos sistemas necesitan realizar un proceso de entrenamiento para cada locutor. El usuario debe proveer algunas muestras de su voz antes de poder usar el sistema (dependencia del locutor). Aunque depende de la tarea, en general, cuanto mayor es el tamaño del vocabulario mayores son las dificultades para el reconocedor. Este aspecto está fuertemente interrelacionado con las similitudes fonéticas que presenta el conjunto de las palabras del vocabulario.

El habla espontanea, por su parte contiene o puede contener cambios bruscos y acusados en la prosodia. Al mismo tiempo, las realizaciones de voz pueden ajustarse a secuencias determinadas de palabras con un contexto explícito o en todo caso siguiendo una gramática muy reducida, lo que limita el número de combinaciones posibles entre palabras, o por el contrario acercarse a realizaciones de lenguaje natural.

Una medida usual de la dificultad de la tarea a realizar por un sistema y que combina el tamaño del vocabulario y el modelo de lengua es el que se conoce como confusión. Depende de la media geométrica del número de palabras que pueden seguir a una dada, una vez que se ha aplicado el modelo de lengua correspondiente.

Por último, encontramos algunos factores del entorno como son el ruido presente en el ambiente de trabajo y la calidad de los mecanismos de captura de la voz; desde micrófonos específicos que realizan parte del filtrado del ruido ambiente, hasta micrófonos de características variables según el usuario (teléfono).

3.1.2. Esquema de un reconocedor de voz

La Figura 3.1 muestra el esquema general de los componentes principales de un sistema típico de reconocimiento de voz.

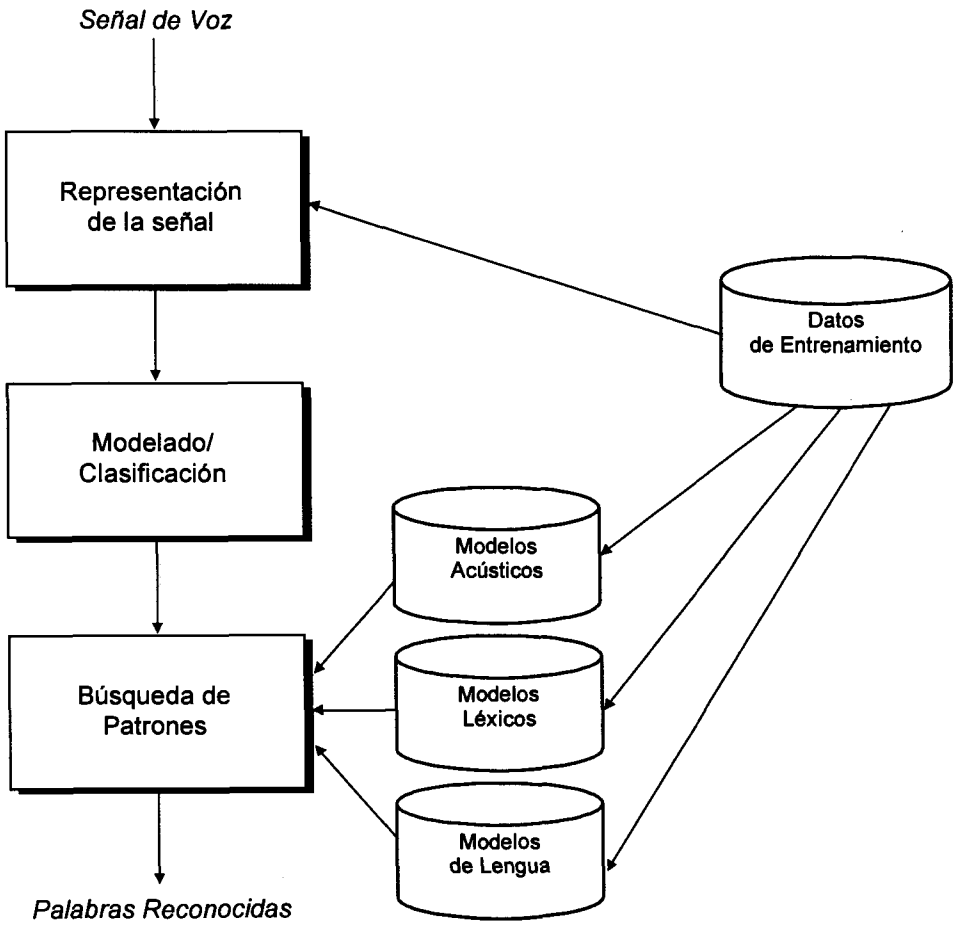

Figura 3.1. Componentes de un sistema de reconocimiento del habla genérico.

Tras obtener los datos de entrada al sistema por medio de la digitalización de la señal de voz, las muestras del paso anterior se transforman en un conjunto de medidas o características útiles. Esta es la fase que llamamos de representación de la señal. La realización de esta tarea se lleva a cabo a un ritmo constante, usualmente de entre 10 y 20 ms y como producto de salida se obtiene una serie de vectores que de alguna manera contienen en esencia las características definitorias de la señal original pero con un tamaño mucho menor.

Posteriormente, las medidas de la etapa anterior se emplean para encontrar la palabra candidato que proporcione el mejor encaje, teniendo en cuenta las restricciones impuestas por los modelos acústico, léxico y de lengua (gramática) disponibles. Esta es la fase de búsqueda de patrones.

Los sistemas de reconocimiento intentan modelar las fuentes de variación descritas antes, de diversas maneras (fase de modelado/clasificación). En el nivel de representación de la señal, los desarrolladores emplean representaciones que enfaticen las características importantes desde un punto de vista perceptual y, por otra parte, tratan de minimizar aquellas otras que son dependientes de determinados hablantes.

La adaptación a un determinado hablante o a las condiciones del entorno de funcionamiento del sistema se realiza a través de un conjunto de procesos de normalización que tienen como finalidad restaurar las características del vector con las medidas a valores lo más próximos posibles a los que se obtendrían en condiciones neutras o a los presentes cuando se entrenó el reconocedor.

El entrenamiento, constituye una etapa previa a la entrada en funcionamiento del conjunto y tiene como finalidad establecer los diferentes conjuntos de modelos empleados durante el proceso de búsqueda:

1. Modelos acústicos que recojan realizaciones dependientes del género del locutor, entonaciones, variantes dialectales, etc.
2. Modelos léxicos que contemplen pronunciaciones alternativas de las palabras, con objeto de permitir que los algoritmos de búsqueda encuentren diferentes caminos.
3. Modelos de lengua que permiten estimar la frecuencia de ocurrencia de determinadas secuencias de palabras.

Además durante el entrenamiento se puede obtener también la información necesaria para ajustar los procesos de modelado y normalización.

3.2. ANÁLISIS ESPECTRAL

3.2.1. Análisis por banco de filtros digitales

El uso de bancos de filtros digitales [PIC93] [ROB98], implementados inicialmente como filtros analógicos, ha sido históricamente la primera aproximación al procesamiento del habla.

Un banco de filtros de paso banda puede entenderse como un modelo sencillo de las etapas iniciales del sistema auditivo humano. El planteamiento puede resumirse en que el banco de filtros intenta descomponer la señal inicial en un conjunto discreto de muestras espectrales que contienen una información similar a la que se presenta en los niveles superiores del sistema auditivo.

Con objeto de aproximarse a la sensibilidad del oído humano, que no tiene una respuesta lineal en frecuencia existen diferentes escalas. Algunos ejemplos expresados de forma analítica, siendo f el valor de frecuencia en Hz, son:

1. Escala de Bark.

$$Bark = 13\,arctan\left(\frac{0.76f}{1000}\right) + 3.5\,atan\left(\frac{f^2}{(7500)^2}\right) \tag{3.2}$$

2. Escala de Mel.

$$m = 2595\log_{10}\left(\frac{1+f}{700}\right) \tag{3.3}$$

Ésta última es la más usual en aplicaciones de tratamiento de la voz.

Un banco de filtros está constituido por un conjunto de filtros cada uno de los cuales retiene la información de una serie determinada de frecuencias del espectro. A su vez cada filtro puede ponderar de manera diferente las frecuencias que quedan bajo su ámbito. Un ejemplo de banco de filtros empleando escalas de Mel y 19 filtros es el de la

Figura 3.2.

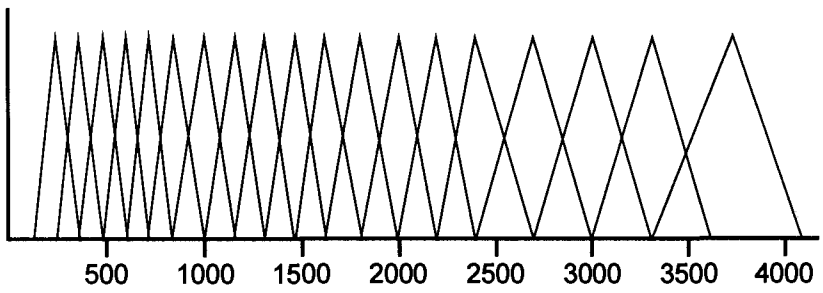

Figura 3.2. Banco de filtros de Holmes.

Este tipo de técnica, generalmente se emplea de manera conjunta con otros métodos como son el cálculo de coeficientes cepstrales.

3.2.2. Transformada discreta de Fourier

La transformada discreta de Fourier o DFT (*Discrete Fourier Transform*) [COO92] [DEL93] [KRA94] [DEL94], se define como:

$$X(k) = \sum_{k=0}^{N-1} x(n)\, e^{-j\left(\frac{2\pi kn}{N}\right)} \quad k = 0,1,2,\ldots,N-1 \tag{3.4}$$

donde N es el número de muestras de la ventana que se va a analizar. La DFT es periódica en la frecuencia con periodo f_s (frecuencia de muestreo) y presenta también como propiedades interesantes su linealidad y la relación existente entre las operaciones de multiplicación/convolución en los dominios temporal y frecuencial.

$$\alpha h(n) + \beta g(n) \Leftrightarrow \alpha H(k) + \beta G(k) \tag{3.5}$$

$$h(n) \otimes g(n) \Leftrightarrow H(k)G(k) \tag{3.6}$$

$$h(n)g(n) \Leftrightarrow H(k) \otimes G(k) \tag{3.7}$$

Por su parte la DFT inversa o IDFT se define como:

$$x(n) = \sum_{k=0}^{N-1} X(k) e^{j\left(\frac{2\pi kn}{N}\right)} \quad n = 0,1,2,\ldots,N-1 \tag{3.8}$$

La motivación del uso de la DFT parte del hecho de la utilidad que tiene descomponer la señal de voz de partida en sus componentes en frecuencia. Para comprender el origen de la expresión de partida nos basaremos en el formalismo de los espacios de Hilbert (espacios infinito-dimensionales) [ALE76], que nos brindan un camino simple y elegante para obtener la expresión básica de traspaso de señales en el dominio del tiempo al dominio de la frecuencia mediante la transformación de Fourier.

Para comenzar, repasaremos algunas propiedades de los espacios n-dimensionales que nos servirán como referencia para entender y emplear sus equivalentes en los espacios de Hilbert.

Sean $f = [f_1, f_2, \ldots, f_n]$ y $g = [g_1, g_2, \ldots, g_n]$ dos vectores n-dimensionales. Definimos las siguientes operaciones:

$$f + g = [f_1 + g_1, f_2 + g_2, \ldots, f_n + g_n] \tag{3.9}$$

$$\lambda f = [\lambda f_1, \lambda f_2, \cdots, \lambda f_n] \quad \lambda \in \Re \tag{3.10}$$

$$\text{Longitud de } f = \sqrt{f_1^2 + f_2^2 + \cdots + f_n^2} = \sqrt{\sum_{i=1}^{n} f_i^2} \tag{3.11}$$

Producto escalar

$$(f, g) = f_1 g_1 + f_2 g_2 + \cdots + f_n g_n = \sum_{i=1}^{n} f_i g_i \tag{3.12}$$

$$(f, g) = \sqrt{f_1^2 + f_2^2 + \cdots + f_n^2}\sqrt{g_1^2 + g_2^2 + \cdots + g_n^2}\cos\alpha =$$

$$= \sqrt{\sum_{i=1}^{n} f_i^2}\sqrt{\sum_{i=1}^{n} g_i^2}\cos\alpha$$

En el caso de que un vector u tenga como longitud la unidad, se puede observar con claridad como el producto escalar de un vector f por u: (f, u) se corresponde con la proyección de f sobre el eje determinado por u.

Como caso particular de suma importancia se encuentra la condición de ortogonalidad entre vectores. Dados dos vectores no nulos f y g, se dice que son ortogonales si su producto escalar es nulo: $(f, g) = 0 \Rightarrow \cos(\alpha)=0 \Rightarrow \alpha=90^o$.

Podemos utilizar las características de los espacios n-dimensionales para trabajar con funciones discretas, para ello basta con considerar las distintas muestras de la función como componentes de un vector multidimensional $f = [y_1, y_2, \ldots, y_n]$.

Los espacios de Hilbert son infinito-dimensionales, por lo que nos encontramos con valores de n tendiendo a infinito y distancia entre componentes tendiendo a cero. En este caso estamos preparados para trabajar con funciones continuas, de hecho, un vector en un espacio de Hilbert se define como una función $f(t) \mid t \in [a,b]$.

En espacios infinito-dimensionales, la suma de vectores y la multiplicación de un

vector por un número se definen como la adición de funciones y la multiplicación de una función por un número. En un espacio de Hilbert donde f es una función tenemos de manera análoga al caso de los espacios n-dimensionales:

$$\text{Longitud } f = \sqrt{\int_a^b f^2(t)\,dt} \tag{3.13}$$

$$\begin{aligned}(f,g) &= \int_a^b f(t)g(t)\,dt \\ (f,g) &= \sqrt{\int_a^b f(t)dt}\sqrt{\int_a^b g(t)dt}\cos\alpha\end{aligned} \tag{3.14}$$

De forma similar a lo que ocurría en los espacios n-dimensionales, en un espacio de Hilbert las funciones cuyo producto escalar es cero son ortogonales. En los espacios n-dimensionales también se cumple la siguiente propiedad: si tomamos m vectores arbitrarios u_i, perpendiculares entre sí y de longitud unidad, todo vector n-dimensional se puede caracterizar mediante sus componentes halladas realizando la proyección del vector sobre cada uno de los n ejes determinados por los vectores u_i:

$$a_k = (f,u_k) \quad k = 1,2,\cdots,m \tag{3.15}$$

En espacios de Hilbert, un sistema de funciones $\phi_1(t)$, $\phi_2(t)$, ... $\phi_n(t)$ es ortogonal si se cumple:

$$\int_a^b \phi_j(t)\phi_k(t)dt = 0 \quad \forall j \neq k \tag{3.16}$$

y es ortonormal si cumple:

$$\int_a^b \phi_i^2(t)dt = 1 \quad i = 1,\cdots,n \tag{3.17}$$

Tomando la sucesión de funciones: 1, $e^{-j\omega t}$, $e^{-j2\omega t}$, $e^{-j3\omega t}$, $e^{-jn\omega t}$, donde $\omega = 2\pi f$ y $f = 1/T$ la frecuencia de la onda de armónico fundamental dado por T, obtenemos la base:

$$\phi_m = e^{-jm\omega t} = \cos(m\omega t) - j\,\text{sen}(m\omega t) \tag{3.18}$$

De esta forma una función $f(t)$ se podrá expresar como:

$$f(t) = a_1\phi_1(t) + a_2\phi_2(t) + \cdots + a_m\phi_m(t) \tag{3.19}$$

Aplicando la expresión del producto escalar y la propiedad de proyección de un vector (o función en este caso) sobre vectores unitarios perpendiculares tenemos:

$$a_m = \int_a^b f(t)\phi_m(t)dt \tag{3.20}$$

La expresión ilustra la descomposición de una señal compleja empleando una base de senos y cosenos de diferentes frecuencias. El producto escalar de la función analizada por cada elemento de la base nos indica la proporción en la que cada frecuencia *(mω)* participa en la composición de la señal original.

Cuando existe una marcada ortogonalidad entre la señal analizada y uno de los elementos de la base, por ejemplo el ϕ_p, significa que la frecuencia determinada por el eje ϕ_p no interviene de forma significativa en la constitución de la señal original. Analíticamente, esta característica se refleja en el término *cos* α de la fórmula del producto escalar. Gráficamente, los términos a_m se corresponden con las proyecciones de la onda analizada sobre cada uno de los ejes determinados por los elementos de la base.

Combinando la anterior expresión y la definición que hicimos de base, los valores a_m se pueden expresar como:

$$a_m = \int_a^b f(t)e^{-jm\omega t}dt \tag{3.21}$$

Por último si trabajamos con N valores discretos $f(n\tau)$ cuyo recorrido se abarca desde un índice $k= 0$ hasta $k= N-1$, esta última ecuación se convierte en la expresión en series de Fourier:

$$a_m = \frac{1}{N}\sum_{n=0}^{N-1} f(n\tau)e^{-jm\frac{2\pi n}{N}} \tag{3.22}$$

Un aspecto importante si queremos usar la DFT con señales de voz es que debemos asumir que al menos en periodos cortos de tiempo se cumple que la señal es

estacionaria. En la realidad esto no es estrictamente así aunque podemos suponerlo. La solución para poder aplicar este método es multiplicar la señal por una función ventana que sea 0 fuera de un determinado rango y reproducir el resultado de forma que tengamos un número de bloques iguales, es decir, que hayamos construido una señal periódica a partir de nuestro fragmento inicial.

La ventana rectangular se define como:

$$w_n = \begin{cases} 1 & 0 \le n \le N \\ 0 & resto \end{cases} \tag{3.23}$$

Sin embargo la utilización de esta ventana trae consigo, que en los puntos de inicio y fin exista una fuerte discontinuidad. Para ello debemos emplear tipos de ventana que tiendan a reducir a 0 los valores de las muestras en los extremos y así reducir el efecto de discontinuidad al mínimo. Aunque existe un buen número de tipos de ventana, la más común en el análisis de la voz es la que se conoce como ventana de Hamming (Figura 3.3):

$$w_n = \begin{cases} 0.54 - 0.46\cos\left(2\pi n/(N-1)\right) & 0 \le n \le N \\ 0 & resto \end{cases} \tag{3.24}$$

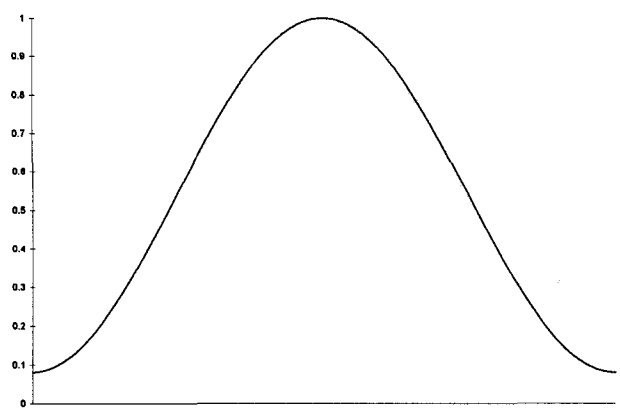

Figura 3.3. Ventana de Hamming.

El efecto de multiplicar los valores de las muestras de voz en el tiempo por los valores de la ventana equivale a convolucionar ambas señales en el dominio de la frecuencia.

Por último, existe un aspecto relevante a la hora de aplicar el método de la DFT y es su complejidad. La complejidad de la DFT es de $O(n^2)$ operaciones y con objeto de acelerar el cálculo de este procedimiento, se emplea habitualmente lo que se conoce como transformada rápida de Fourier o FFT (*Fast Fourier Transform*). Simplemente es una manera eficiente de computar la DFT y su complejidad es de $O(n \log n)$ si n es una potencia de 2.

El algoritmo de "diezmado en el tiempo" de Cooley y Tukey comienza con la expresión de la DFT.

$$X_1(k) = \sum_{n=0}^{N-1} x_n e^{\frac{-i2\pi nk}{N}}, \quad k = 0,\ldots,N-1 \tag{3.25}$$

El objetivo es alcanzar una expresión recursiva que permita obtener el resultado deseado dividiendo el problema en subproblemas más sencillos.

$$X_1(k) = \sum_{n=0}^{N/2-1} x_{2n} e^{\frac{-i2\pi(2n)k}{N}} + \sum_{n=0}^{N/2-1} x_{2n+1} e^{\frac{-i2\pi(2n+1)k}{N}} = \tag{3.26}$$

$$= \sum_{n=0}^{N/2-1} x_{2n} e^{\frac{-i2\pi(2n)k}{N}} + e^{\frac{-i2\pi k}{N}} \sum_{n=0}^{N/2-1} x_{2n+1} e^{\frac{-i2\pi(2n)k}{N}} = \tag{3.27}$$

$$= \sum_{n=0}^{N/2-1} x_{2n} e^{\frac{-i2\pi nk}{(N/2)}} + e^{\frac{-i2\pi k}{N}} \sum_{n=0}^{N/2-1} x_{2n+1} e^{\frac{-i2\pi nk}{N/2}} = \tag{3.28}$$

$$= X_{11}(k) + e^{\frac{-i2\pi k}{N}} X_{12}(k) \tag{3.29}$$

Así la transformada inicial se ha dividido en dos transformadas que operan sobre las porciones par e impar de los datos. Aunque la complejidad de cada porción es del tipo $O(n^2)$ el número de puntos es ahora la mitad. La recursión puede continuarse hasta el caso trivial de tener solamente un punto, que nos lleva al $O(n \log n)$ comentado.

3.2.3. Predicción lineal

El método de predicción lineal o LP (*Linear Prediction*) [ITA70] [MAK75] [HAY96] [DEL93] es históricamente uno de los métodos más importantes para el análisis de la voz. Su fundamento se basa en establecer un modelo de filtro del tipo todo polo, para la fuente de sonido.

La principal motivación del modelo todo polo viene dada porque permite describir la función de transferencia de un tubo, que sin pérdidas estuviese formado por diferentes secciones. Ésta es a su vez una aproximación razonable al habla producida por la excitación del tracto vocal causada por el conjunto de pulsos glotales. Sin embargo, entre las objeciones que pueden establecerse a esta asociación encontramos:

1. Los pulsos glotales no tienen una estructura espectral plana.
2. El tracto vocal no está compuesto de cilindros.
3. La cavidad nasal constituye un pasaje adicional.
4. Algunos sonidos se generan cerca de los labios como algunos sonidos fricativos sordos.

Sin embargo, con un número suficiente de parámetros el modelo de predicción lineal puede constituir una aproximación adecuada a la estructura espectral de todo tipo de sonidos.

El método de predicción lineal recibe este nombre porque pretende extrapolar el valor de la siguiente muestra de voz $x(n)$ como la suma ponderada de muestras pasadas $x(n-1)$, $x(n-2)$, ..., $x(n-K)$:

$$\tilde{x}(n) = \sum_{i=1}^{K} a_i x(n-i) \tag{3.30}$$

Para ello se debe realizar el cálculo de los coeficientes a_i minimizando alguna función de error E, concretamente de mínimos cuadrados, sobre una ventana de tamaño N.

$$E = \sum_{n=0}^{N-1} e^2(n) = \sum_{n=0}^{N-1} \left(x(n) - \sum_{i=1}^{K} a_i x(n-i) \right)^2 \quad 0 \le n \le N-1 \tag{3.31}$$

El mínimo valor de E se obtendrá haciendo cero la derivada con respecto a cada uno de los a_j. con $1 \le j \le K$.

$$\frac{\partial E}{\partial a_j} = 0 \quad 1 \le j \le K \tag{3.32}$$

$$\frac{\partial E}{\partial a_j} = \frac{\partial}{\partial a_j} \sum_n \left[x(n) - \sum_{i=1}^{K} a_i x(n-i) \right]^2 = 0 \tag{3.33}$$

$$\frac{\partial E}{\partial a_j} = \sum_n \frac{\partial}{\partial a_j} \left[x(n) - \sum_{i=1}^{K} a_i x(n-i) \right]^2 = 0 \tag{3.34}$$

$$\frac{\partial E}{\partial a_j} = \sum_n 2 \frac{\partial}{\partial a_j} \left[x(n) - \sum_{i=1}^{K} a_i x(n-i) \right] * \left[x(n) - \sum_{i=1}^{K} a_i x(n-i) \right] = 0 \tag{3.35}$$

$$\frac{\partial E}{\partial a_j} = -2 \sum_n x(n-j) \left[x(n) - \sum_{i=1}^{K} a_i x(n-i) \right] = -2 \sum_n x(n-j) e(n) = 0 \tag{3.36}$$

$$\frac{\partial E}{\partial a_j} = \sum_n x(n-j) e(n) = 0 \quad 1 \le j \le K \tag{3.37}$$

Definiendo c_{ji}, la expresión puede quedar:

$$c_{ji} = \sum_n x(n-j) x(n-i) \tag{3.38}$$

$$\sum_n x(n-j) \left[x(n) - \sum_{i=1}^{K} a_i x(n-i) \right] = \tag{3.39}$$

$$= \sum_n x(n-j) x(n) - \sum_{i=1}^{K} a_i \sum_n x(n-j) x(n-i) = \tag{3.40}$$

$$= c_{j0} - \sum_{i=1}^{K} a_i c_{ji} \quad 1 \le i \le K \quad 1 \le j \le K \tag{3.41}$$

Esto constituye un conjunto de ecuaciones normales que permiten el cálculo de los coeficientes a_i:

$$\begin{aligned} & c_{10} - a_1 c_{11} - a_1 c_{12} - \ldots - a_1 c_{1K} = 0 \\ & \qquad \ldots \\ & c_{K0} - a_1 c_{K1} - a_1 c_{K2} - \ldots - a_1 c_{KK} = 0 \end{aligned} \tag{3.42}$$

$$\begin{bmatrix} c_{10} \\ \ldots \\ c_{K0} \end{bmatrix} = \begin{bmatrix} c_1^T \\ \ldots \\ c_K^T \end{bmatrix} a \tag{3.43}$$

$$q = Ca \tag{3.44}$$

Las ecuaciones normales tienen una solución intuitiva mediante la inversión matricial (método de la covarianza):

$$a = C^{-1} q \tag{3.45}$$

Una manera sencilla de definir los límites de n en el sumatorio, que permite obtener los valores de los coeficiente c_{ij} de la matriz C, es suponer que el valor de las muestras de voz se anula fuera del intervalo $0 \le n \le N-1$. Este método se llama de autocorrelación y es equivalente a aplicar una ventana rectangular en el intervalo considerado, de esta manera:

$$\sum_n x(n-i)\,x(n-j) = \sum_n x(n)\,x(n+|i-j|) = r_{|i-j|} \tag{3.46}$$

con lo que $C_{ij} = C_{ji} = r_{|i-j|}$, donde los $r_{|i-j|}$ son los coeficientes de correlación de la matriz de autocorrelación. Esta matriz es simétrica y todos los elementos de la diagonal son iguales. Además podría ser generada a partir de un único vector. Esto la convierte en una matriz tipo Toeplitz:

$$\sum_{i=1}^{K} r_n(|j-i|)\, a_i = r_n(j) \quad i \le j \le K \tag{3.47}$$

Expresado en forma matricial:

$$\begin{bmatrix} r_n(0) & r_n(1) & r_n(2) & \dots & r_n(K-1) \\ r_n(1) & r_n(0) & r_n(1) & \dots & r_n(K-2) \\ r_n(2) & r_n(1) & r_n(0) & \dots & r_n(K-3) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ r_n(K-1) & r_n(K-2) & r_n(K-3) & \dots & r_n(0) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ \vdots \\ a_K \end{bmatrix} = \begin{bmatrix} r_n(1) \\ r_n(2) \\ r_n(3) \\ \vdots \\ r_n(K) \end{bmatrix} \tag{3.48}$$

Existen diferentes métodos para invertir este tipo de matrices. Uno de ellos es el algoritmo de Levinson-Durbin. Denotamos como $a_j^{(i)}$ al parámetro j de la predicción lineal en la iteración i y como $L^{(i)}$ a la energía residual en dicha iteración. Los parámetros h_i se conocen como coeficientes de reflexión o de correlación parcial PARCOR (PARtial CORrelation).

$$L^{(0)} = r(0) \tag{3.49}$$

$$h_i = a_i^{(i)} = \frac{r(i) - \sum_{j=1}^{i-1} a_j^{(i-1)} r(|i-j|)}{L^{(i-1)}} \quad 1 \le i \le K \tag{3.50}$$

$$h_i = -a_i^{(i)} \tag{3.51}$$

$$a_j^{(i)} = a_j^{(i-1)} + h_i a_{i-j}^{(i-1)} \tag{3.52}$$

$$L^{(i)} = \left(1 - h_i^2\right) L^{(i-1)} \tag{3.53}$$

Partiendo de las ecuaciones anteriores y de:

$$e(n) = x(n) - \sum_{i=1}^{k} a_i(n)\, x(n-i) \tag{3.54}$$

$$f_k(n) = x(n-k) - \sum_{i=0}^{k-1} a_{k-i}(n-k)\, x(n-i) \tag{3.55}$$

obtenemos las expresiones del filtro predictor y postdictor del error, respectivamente:

$$e_k(n) = e_{k-1}(n) + h_k\, f_{k-1}(n-1) \tag{3.56}$$

$$f_k(n) = f_{k-1}(n-1) + h_k\, e_k(n) \tag{3.57}$$

Empleando la definición de $a_i^{(i)}$ del método Levinson-Durbin se puede obtener una expresión alternativa para los coeficientes de reflexión que utilice solamente los errores directo e inverso:

$$h_{dk} = -\frac{\sum_n e_{k-1}(n)\, f_{k-1}(n-1)}{\sum_n f_{k-1}^2(n-1)} \tag{3.58}$$

$$h_{ik} = -\frac{\sum_n e_{k-1}(n)\, f_{k-1}(n-1)}{\sum_n e_{k-1}^2(n-1)} \tag{3.59}$$

Donde h_{dk} es el coeficiente directo y h_{ik} es el coeficiente inverso. Si el proceso es estacionario, ambos coeficientes tomarán el mismo valor y las normas de los dos vectores error serán las mismas. Si no es así, se puede utilizar el estimador que se obtiene de realizar la media geométrica de h_{dk} y h_{ik}, también conocido como método de Itakura:

$$h_{lk} = -\frac{\sum_n e_{k-1}(n)\, f_{k-1}(n-1)}{\sqrt{\sum_n e_{k-1}^2(n) \sum_n f_{k-1}^2(n-1)}} \tag{3.60}$$

Con estas expresiones se pueden calcular los coeficientes de forma adaptativa, para ello se toman estimadores de los errores, los cuales se pueden actualizar en cualquier punto espacial o temporal realizando un sencillo cálculo.

Si el numerador y denominador de la anterior expresión se estiman mediante:

$$T_{k-1}(n)=\mu T_{k-1}(n-1)+2\, e_{k-1}(n) f_{k-1}(n-1) \tag{3.61}$$

$$L_{k-1}(n)=\mu\ L_{k-1}(n-1)+e_{k-1}^2(n)+f_{k-1}^2(n-1) \quad 0\le\mu\le 1 \tag{3.62}$$

el coeficiente queda entonces como (Método de Burg):

$$h_k(n+1)=-\frac{T_{k-1}(n)}{L_{k-1}(n)} \quad 1\le k\le K \tag{3.63}$$

Para poder determinar de manera recursiva el valor del coeficiente, variamos la expresión a:

$$h_k(n+1)=h_k(n)+\delta_k(n) \tag{3.64}$$

$$\begin{aligned}\delta_k(n)=h_k(n+1)-h_k(n)=\frac{T_{k-1}(n)}{L_{k-1}(n)}-\frac{T_{k-1}(n-1)}{L_{k-1}(n-1)}=\\ =-\frac{1}{L_{k-1}(n)}\left[T_{k-1}(n)-\frac{T_{k-1}(n-1)\ L_{k-1}(n)}{L_{k-1}(n-1)}\right]=\end{aligned} \tag{3.65}$$

$$\begin{aligned}=-\frac{1}{L_{k-1}(n)}\Big\{\left[\mu T_{k-1}(n-1)+2\,e_{k-1}(n)f_{k-1}(n-1)\right]-\\ -\frac{T_{k-1}(n-1)}{L_{k-1}(n-1)}\left[\mu L_{k-1}(n-1)+e_{k-1}^2(n)+f_{k-1}^2(n-1)\right]\Big\}=\end{aligned} \tag{3.66}$$

$$=-\frac{1}{L_{k-1}(n)}\left\{2\,e_{k-1}(n)f_{k-1}(n-1)+h_k(n)\left[e_{k-1}^2(n)+f_{k-1}^2(n-1)\right]\right\}= \tag{3.67}$$

$$\begin{aligned}=-\frac{1}{L_{k-1}(n)}\Big\{\,e_{k-1}(n)\left[f_{k-1}(n-1)+h_k(n)e_{k-1}(n)\right]+\\ +f_{k-1}(n-1)\left[e_{k-1}(n)+h_k(n)f_{k-1}(n-1)\right]\Big\}=\end{aligned} \tag{3.68}$$

$$=-\frac{1}{L_{k-1}(n)}\left[e_{k-1}(n)f_k(n)+f_{k-1}(n-1)e_k(n)\right]= \tag{3.69}$$

Con lo que la expresión recursiva buscada queda como:

$$h_k(n+1) = h_k(n) - \frac{1}{L_{k-1}(n)}\left[e_{k-1}(n)f_k(n) + f_{k-1}(n-1)e_k(n)\right] \tag{3.70}$$

Ahora podemos escribir el algoritmo celosía adaptativa de gradiente o GAL (*Gradient Adaptive Lattice*) como:

1. Inicio:

Para $1 \leq k \leq K$

$$h_k(0) = 0$$

$$L_{k-1}(0) = \text{Constante próxima a } 0$$

2. Bucle principal:

Para $0 \leq n \leq N-1$

$$e_0(n) = f_0(n) = x(n)$$

Para $1 \leq k \leq K$

$$h_k(n+1) = h_k(n) - \frac{1}{L_{k-1}(n)} *$$

$$*\left[e_{k-1}(n)f_k(n) + f_{k-1}(n-1)e_k(n)\right]$$

$$e_k(n+1) = e_{k-1}(n+1) + h_k(n+1)f_{k-1}(n)$$

$$f_k(n+1) = f_{k-1}(n) + h_k(n+1)e_k(n+1)$$

$$L_{k-1}(n+1) = \mu\, L_{k-1}(n) + e_{k-1}^2(n+1) + f_{k-1}^2(n)$$

Este algoritmo tiene la propiedad de que los grados de convergencia de los coeficientes de reflexión son independientes de la dispersión de los autovalores de la matriz de correlación.

Cada etapa de la celosía intenta adaptarse a la componente espectral más prominente de la señal muestra a muestra. Además es importante la utilidad del parámetro de ajuste μ, puesto que una disminución en éste hace que el algoritmo no responda tan rápidamente a las condiciones externas, lo cual es útil en circunstancias de ruido.

A diferencia del algoritmo de Levinson-Durbin cuyos coeficientes se pueden pasar directamente a la función de transferencia, los valores obtenidos tras el procesamiento

de la celosía de gradiente adaptativo requieren de una fase de adaptación.

Los valores que tenemos son los coeficientes de pivotaje $a_i^{(i)}$, para obtener los $a_i^{(j)}$ buscados, aplicamos la conversión descrita en el algoritmo de Levinson-Durbin.

1. Inicio:

Para $0 \leq j \leq K$

$$e_i^{(j)} = 0$$

2. Bucle principal:

Para $1 \leq k \leq K$

$$a_i^0 = -1$$

Para $0 \leq i \leq k$

$$c_i = a_i^i + a_{k-1}^{k-1} a_i^{k-i}$$

Para $0 \leq i \leq K$

$$a_i^i = c_i$$

Una vez que se ha obtenido el conjunto de coeficientes LPC de orden K, $[1, -a_{1k}, \ldots, -a_{kk}]^T$, el último paso es obtener las características espectrales de la señal de la cual se obtuvieron los coeficientes. La función de transferencia inversa $F_k(z)$ permite obtener el espectrograma deseado:

$$F_k(z) = \frac{1}{1 - \sum_{i=1}^{k} a_{ik} z^{-k}} \tag{3.71}$$

de la cual evaluaremos su módulo sobre el círculo unidad:

$$\left| F_k\left(z = e^{j\omega\tau}\right) \right| = \frac{1}{\left| 1 - \sum_{i=1}^{k} a_{ik} e^{-ji\omega\tau} \right|} \tag{3.72}$$

Dado que en sistema discreto no podemos manejar funciones continuas, sino vectores discretos representantes de las mismas, recurrimos a muestrear dicha función de transferencia en intervalos discretos de la variable ω representante de la frecuencia

angular, dados por múltiplos del ángulo mínimo Ω o resolución en frecuencia:

$$\Omega = \frac{\pi}{M\tau} \tag{3.73}$$

$$\omega = m\Omega \tag{3.74}$$

siendo M el número de muestras que deseamos para la función de transferencia y m el índice discreto de frecuencia.

La evaluación del módulo de la función de transferencia, deberá hacerse en el dominio discreto, teniendo en cuenta:

$$e^{-ji\omega\tau} = \cos(im\Omega\tau) - j\,\mathrm{sen}(im\Omega\tau) = \cos\left(i\frac{m}{M}\pi\right) - j\,\mathrm{sen}\left(i\frac{m}{M}\pi\right) \tag{3.75}$$

De este modo, la evaluación del módulo de la función de transferencia, se hará como sigue:

$$|F_k(m)| = \frac{1}{\sqrt{\left[1 - \sum_{i=1}^{k} a_{ik}\cos\left(i\frac{m}{M}\pi\right)\right]^2 + \left[\sum_{i=1}^{k} a_{ik}\,\mathrm{sen}\left(i\frac{m}{M}\pi\right)\right]^2}} \tag{3.76}$$

3.2.4. Cepstrum

Desde la introducción en los primeros años de la década de los 70, de las técnicas homomórficas de procesado de señal [OPP75] [PIC93], su importancia dentro del campo del reconocimiento de voz ha sido muy grande. Los sistemas homomórficos son una clase de sistemas no lineales que obedecen a un principio de superposición. De éstos, los sistemas lineales constituyen un caso especial. La motivación para realizar un procesado homomórfico del habla viene resumida en la Figura 3.4:

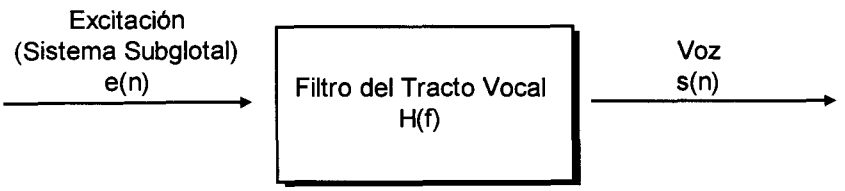

Figura 3.4. Las técnicas homomórficas pueden servir para separar la acción del tracto vocal (filtro lineal variable en el tiempo) de la señal de excitación.

En el procesado de señal de voz, el sistema homomórfico que estamos buscando debe tener la siguiente propiedad:

$$D\left[\left[x_1(n)\right]^{\alpha} \bullet \left[x_2(n)\right]^{\beta}\right] = \alpha D\left[x_1(n)\right] + \beta D\left[x_2(n)\right] \tag{3.77}$$

Éste es el tipo de operación de superposición con respecto a la multiplicación, exponenciación y suma. La función logaritmo, por supuesto, obedece dicha propiedad.

La señal de voz *s(n)* se descompone en una parte de excitación *e(n)* y en un filtro lineal $H(e^{i\theta})$. Así en el dominio frecuencia tenemos:

$$S\left(e^{i\theta}\right) = H\left(e^{i\theta}\right)E\left(e^{i\theta}\right) \tag{3.78}$$

Tomando logaritmos (complejos) a ambos lados de la expresión tenemos:

$$\log\left(S\left(e^{i\theta}\right)\right) = \log\left(H\left(e^{i\theta}\right)\right) + \log\left(E\left(e^{i\theta}\right)\right) \tag{3.79}$$

Sin embargo, para la mayoría de las aplicaciones de voz sólo necesitamos la

amplitud espectral, con lo que la ecuación quedaría:

$$\log\left(\left|S\left(e^{i\theta}\right)\right|\right) = \log\left(\left|H\left(e^{i\theta}\right)\right|\right) + \log\left(\left|E\left(e^{i\theta}\right)\right|\right) \tag{3.80}$$

En el dominio logarítmico, las dos componentes anteriores pueden separarse empleando técnicas convencionales de procesamiento de señal.

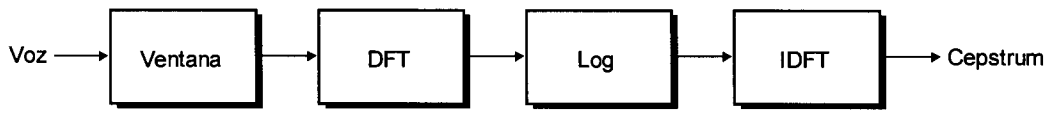

Figura 3.5. Análisis cepstral partiendo de la transformada discreta de Fourier.

Para computar el valor del cepstrum (Figura 3.5), primero se deben calcular las magnitudes espectrales logarítmicas, promediadas si se desea, a través de la suma ponderada de grupos de valores adyacentes agrupados en bandas. A continuación se halla la transformada inversa de Fourier del espectro logarítmico calculado previamente:

$$c(n) = \frac{1}{N_s}\sum_{k=0}^{N_s-1} \log_{10}\left|S_{med}(k)\right| e^{j\frac{2\pi}{N_s}kn} \qquad 0 \le n \le N_s - 1 \tag{3.81}$$

En este caso, los valores *c(n)* se conocen como coeficientes cepstrales derivados de la Transformada de Fourier. N_s es el número de puntos con los que se calculó la DFT.

Esta ecuación también se conoce como la inversa de la DFT del espectro logarítmico. Puede ser convenientemente simplificada teniendo en cuenta que el espectro logaritmo es una función real simétrica:

$$c(n) = \frac{2}{N_s}\sum_{k=1}^{N_s} S_{med}\left(I(k)\right)\cos\left(\frac{2\pi}{N_s}kn\right) \tag{3.82}$$

Lo habitual es usar solamente los primeros términos (n≤ 20). *I(k)* representa una función que traduce la posición de un valor en frecuencia al intervalo donde esté contenido.

Por último, es posible a la hora de calcular un coeficiente cepstral, emplear bandas

definidas según escalas de Mel. Este tipo de parámetros se conoce como coeficientes cepstrales con frecuencia en escalas de Mel o MFCC (*Mel Frequency Cepstral Coefficients*) [DAV80].

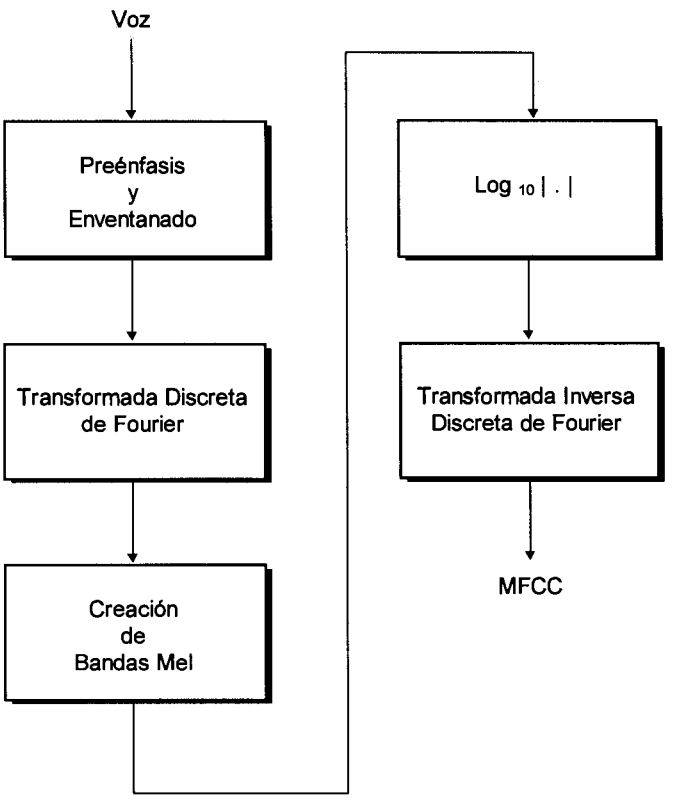

Figura 3.6. Esquema de parametrización para la obtención de MFCC.

Este proceso (Figura 3.6), consiste en tomar los valores en frecuencia obtenidos a través de la DFT y ponderarlos por la serie de filtros definidos en un determinado banco de filtros. Este banco de filtros está constituido por un conjunto de filtros triangulares distribuidos linealmente en una escala de Mel. Posteriormente se emplea la función logaritmo sobre las magnitudes calculadas para cada filtro y finalmente se aplica la transformada discreta del coseno o DCT (*Discrete Cosine Transform*) para obtener el número deseado de coeficientes:

$$MFCC(i) = \sum_{j=1}^{P} m_j \cos\left(\frac{\pi i}{P}(j-0.5)\right) \quad 1 \le i \le K \tag{3.83}$$

donde P es el número de filtros o bandas, K es el número de coeficientes

cepstrales y m_j es el logaritmo de la salida del j-ésimo filtro del banco de filtros.

Partiendo del análisis de predicción lineal también es posible obtener la expresión de los coeficientes cepstrales asociados:

$$c(0) = \log(1) = 0 \tag{3.84}$$

$$c(i) = -a(i) - \sum_{j=1}^{i-1}\left(1 - \frac{j}{i}\right)a(j)c(i-j) \quad 1 \le i \le N_c \tag{3.85}$$

Los coeficientes *a(i)* son los calculados por el método de predicción lineal. N_c es el número de coeficientes cepstrales.

Una transformación usual sobre este tipo de coeficientes es lo que se conoce como coeficientes cepstrales delta o coeficientes delta cepstrum. La expresión que permite obtener estos últimos es:

$$\Delta c_j(i) = \frac{1}{2T+1}\sum_{k=-T}^{T} k \cdot c_{j+k}(i) \tag{3.86}$$

siendo *j* el número de trama sobre la que se halla los coeficientes delta, e *i* el número de coeficiente cepstral. El valor *T* modela la porción de tramas circundantes que se toman en el cálculo de los coeficientes delta.

3.2.5. Predicción lineal perceptual

La técnica de predicción lineal perceptual o PLP (*Perceptual Linear Prediction*) [HER90], es en esencia una combinación de las técnicas de la transformada discreta de Fourier y de predicción lineal como puede verse en la Figura 3.7.

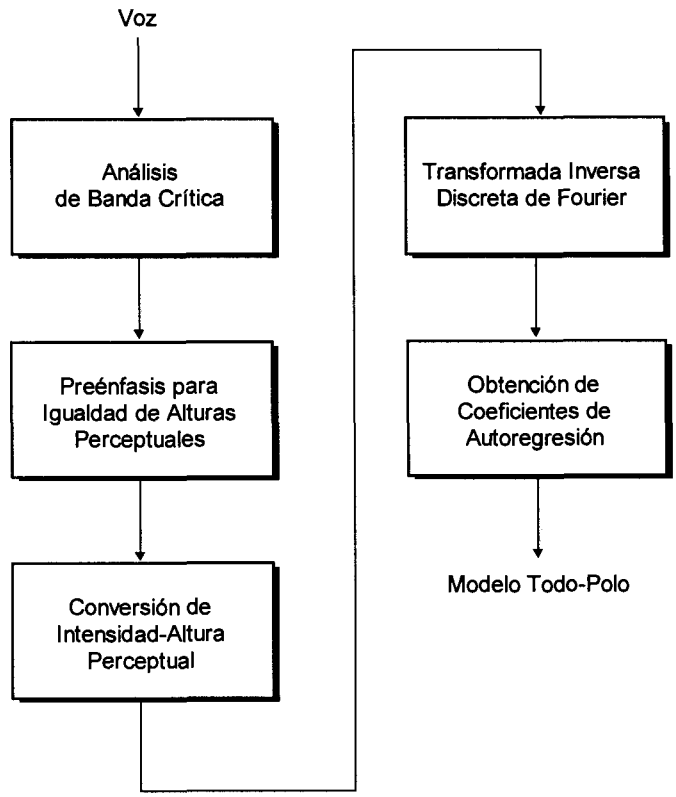

Figura 3.7. Predicción Lineal Perceptual (PLP).

Para obtener el análisis de banda crítica se utiliza primeramente la transformada discreta de Fourier con una ventana de Hamming de 20 ms. Posteriormente se calcula el espectro de potencia y se transfiere a una escala de Bark empleando la siguiente aproximación:

$$\Omega(\omega) = 6\ln\left(\omega/1200\pi + \sqrt{(\omega/1200\pi)^2 + 1}\right) \tag{3.87}$$

A su vez la curva de banda crítica viene dada por:

$$\Psi(\omega) = \begin{cases} 0 & \Omega < -1.3 \\ 10^{2.5(\Omega+0.5)} & -1.3 \le \Omega < -0.5 \\ 1 & -0.5 \le \Omega < 0.5 \\ 10^{-1.0(\Omega-0.5)} & 0.3 \le \Omega < 2.5 \\ 0 & \Omega > 2.5 \end{cases} \tag{3.88}$$

El espectro con escala de Bark se convoluciona con el espectro de potencia del filtro de banda crítica. Esto simula la resolución espectral del oído que es aproximadamente constante en la escala de Bark.

$$\theta(\Omega_i) = \sum_{\Omega=-1.3}^{2.5} P(\Omega - \Omega_i)\Psi(\Omega) \tag{3.89}$$

La operación de convolución reduce la resolución espectral respecto a la inicial $P(\omega)$. El nuevo espectro basado en escalas de Bark suavizadas se muestrea ahora a intervalos de aproximadamente 1 Bark. Así para cubrir el espacio entre 0- 5 kHz (0-16.9 Bark) se requieren 18 muestras espectrales de $\theta[\Omega(\omega)]$.

El segundo paso, consistente en la igualación de las alturas perceptuales tiene su origen en la necesidad de compensar la diferente percepción de alturas sonoras para diferentes frecuencias:

$$\Xi[\Omega(\omega)] = E(\omega)\Theta[\Omega(\omega)] \tag{3.90}$$

$$E(\omega) = \frac{\left(\omega^2 + 56.8x10^6\right)\omega^4}{\left(\omega^2 + 6.3x10^6\right)\left(\omega^2 + 0.38x10^9\right)\left(\omega^6 + 9.58x10^{26}\right)} \tag{3.91}$$

La última etapa previa a la inversa de la transformada discreta de Fourier (IDFT en su siglas inglesas), establece la relación entre altura de sonido percibida e intensidad sonora:

$$\Theta(\omega) = \Xi(\omega)^{1/3} \tag{3.92}$$

Tras la IDFT se calculan los coeficientes de autorregresión de un modelo todo polo. Adicionalmente se pueden calcular a partir de éstos los coeficientes cepstrales.

Como extensión de la técnica anteriormente descrita, encontramos el método RASTA-PLP (RelAtive SpecTrAl) [HER92], [HER94]. La motivación de este complemento viene dada por el intento de robustecer el mecanismo del algoritmo frente a distorsiones lineales en el espectro, por ejemplo debidas al canal de comunicación. Una extensión del algoritmo RASTA es la conocida como J-RASTA [KOE94], que puede también compensar el ruido cuando la relación señal/ruido es baja.

El método RASTA-PLP, incluye además de las etapas propias del PLP las siguientes:

1. Cálculo del logaritmo del espectro de banda crítica tras el cálculo de éste.
2. Cálculo de las derivadas temporales del paso anterior a lo largo de 5 puntos.
3. Reintegrado de la derivada temporal del espectro logarítmico de banda crítica, usando un filtro IIR de primer orden. La posición del polo se establece con un valor de 0.98.
4. Tras calcular la relación entre intensidad-altura perceptual se toma el inverso del logaritmo (función exponencial).

Si la derivada del paso (2) se tomase simplemente como primera diferencia y la operación de reintegrado se hiciera de forma completa (con z= 1.0), entonces el resultado sería equivalente a la sustracción en el espectro logarítmico de la primera trama de análisis frente a cada nueva trama. Sin embargo en este caso, el proceso completo de derivación-reintegración es equivalente a aplicar a cada canal de frecuencia un filtro de banda de paso del tipo IIR, con función de transferencia:

$$H(z) = 0.1 \cdot \frac{2 + z^{-1} - z^{-3} - 2z^{-4}}{z^{-4}\left(1 - 0.98z^{-1}\right)} \tag{3.93}$$

3.3. CUANTIFICACIÓN VECTORIAL

La cuantificación vectorial [MAK85], [GRA84] es una técnica que permite comprimir la información proporcionada por las plantillas espectrales obtenidas en la fase anterior y de paso reducir la dispersión estadística presente en los datos.

La principal motivación para realizar este proceso viene determinada por el modelo de producción de la voz. Si el número de sonidos elementales que se pueden producir físicamente es limitado en una lengua, debemos poder ser capaces de encontrar la transformación que convierta nuestro conjunto de vectores de valores continuos en un nuevo conjunto con un número de vectores finito. En el caso más extremo los vectores pasan a ser elementos individuales.

La gran ventaja de la cuantificación vectorial es que la separación de los vectores en clases no sigue ninguna distribución o función de densidad de probabilidad prefijada.

Un cuantificador vectorial se compone de dos elementos principales: una matriz de transformación de dimensiones N_{vq} x N_y, que se conoce como libro de códigos y una medida de distancia. Las dimensiones de la matriz N_{vq} e N_y se corresponden con el número de vectores del conjunto destino y con la dimensión de los vectores que tenemos en el origen, respectivamente. La distancia permitirá elegir el vector del conjunto final que se encuentre más cerca del vector del que disponemos inicialmente. La medida de distancia que usualmente se emplea es la distancia Euclídea.

En la práctica, sin embargo, la distancia Euclídea se pondera por el valor de la varianza de cada una de las componentes del vector de parámetros de entrada. Las varianzas se calculan con el mismo juego de datos que sirve para crear los libros de códigos. La distancia así calculada recibe el nombre de distancia de Mahalanobis:

$$D(\bar{x},\bar{y}) = \sqrt{\sum_{k=1}^{Ny} C_k (x_k - y_k)^2} \tag{3.94}$$

Para calcular los libros de códigos se pueden emplear diversos algoritmos de los

cuales el más popular es el de las k-medias (*k-means clustering*). Este algoritmo parte de k conjuntos de vectores, es de tipo iterativo y se ejecuta en dos fases:

1. Reagrupar todos los vectores de entrada en k particiones (dependiendo de su cercanía a los centros de clase).
2. Recálculo de los nuevos centroides de clase tomando para ello los vectores que en paso anterior se asignaron a esa clase (Figura 3.8).

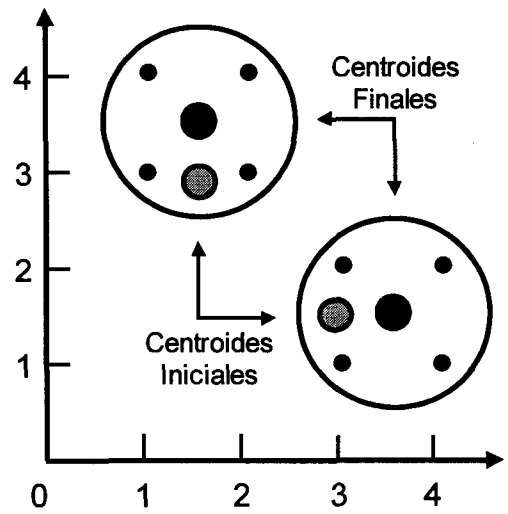

Figura 3.8. Reorganización de los centroides de clase.

Este algoritmo es de tipo iterativo y no garantiza la convergencia a la solución óptima. Sin embargo, aplicado a la voz, la solución que proporciona es suficientemente robusta. La gran ventaja en la utilización de cuantificador vectorial es el ahorro que supone transformar un vector n-dimensional en un valor entero (número de índice de clase).

Otro algoritmo, que permite descomponer sucesivamente el conjunto de los vectores de la entrada en una serie de clases de acuerdo a medidas de distorsión de éstas, es el algoritmo LBG (*Linde, Buzo, Gray*) [LIN80]. El resultado de este algoritmo puede utilizarse como partición inicial del algoritmo de las k-medias.

Otros algoritmos que sirven para realizar este proceso son el LVQ (*Learning*

Vector Quantizer) [KOH90] y el algoritmo de cuantificación vectorial borroso (*fuzzy vector quantization*) [TSE87].

En la práctica, se puede emplear más de un libro de códigos con objeto de poder separar cada grupo de características del vector de parámetros. Así es posible encontrar un libro para los rasgos estáticos y otro para los dinámicos [GUP87]. También puede extenderse a un tercer libro en el que se incluya de forma separada, la información de energía y energía dinámica[LEE90a], [LEE90b].

Por último, es posible encontrar simultáneamente diversos esquemas de cuantificación [PEI95], [PEI96]. Aquí se emplean diversos libros de códigos para un sistema de modelos de Markov semicontinuos. La cuantificación vectorial se realiza empleando el algoritmo LBG junto con un esquema de actualización cercano al modelo LVQ.

3.4. MÉTODOS DE ENCAJE DE PATRONES

3.4.1. Distorsión dinámica temporal

El método de distorsión dinámica temporal o DTW (*Dynamic Time Warping*) [SAK78], [SAK92] es uno de los algoritmos aplicados en reconocimiento de voz [RAB81] más antiguos e importantes, aunque en la actualidad ha cedido paso a otros procedimientos como son los modelos ocultos de Markov. Si bien esta técnica aún se continúa utilizando, tiene un número de limitaciones que restringen su uso a sistemas con vocabularios pequeños pues en sistemas de mayor tamaño, el número de plantillas a generar y el coste computacional de las búsquedas es intratable.

La manera más fácil de reconocer una palabra aislada pasa por compararla con un conjunto de plantillas previamente almacenadas y determinar cual es la que proporciona un mejor encaje. Sin embargo, este objetivo se complica por dos factores: el primero es que la duración de la palabra no tiene que ser la misma que la de las plantillas y el segundo es que el ritmo con el que pronuncia esa palabra no tiene por que ser constante. Resumiendo, el alineamiento óptimo, entre plantillas almacenadas y las plantillas producidas en un determinado momento, puede ser no lineal.

Sean $X= (x_1, x_2, ..., x_I)$ e $Y= (y_1, y_2, ..., y_J)$ dos patrones de voz aislados (ej. palabras). La disparidad promedio entre X e Y, $D(X,Y)$, se basa en alguna medida de distancia entre los vectores x_i e y_j que denotaremos como $d(i,j)$.

La distorsión temporal no lineal de los patrones X e Y puede representarse por un camino $\{ P(k)= (m(k), n(k)), k= 1, K \}$ en el plano (i, j) definida por las dos secuencias de vectores X e Y. La (Figura 3.9) muestra un ejemplo de dicho camino de normalización:

La disparidad entre las tramas X e Y a lo largo de un camino P viene dada por:

$$D_P(X,Y) = \sum_{k=1}^{K} d(P(k)) \cdot \frac{w(k)}{N(w)} \tag{3.95}$$

donde $w(k)$ es un peso y $N(w)$ es el factor de normalización. Existe un número de posibles caminos $P(k)$, que corresponden a diferentes funciones de distorsión para los patrones de voz. El objetivo será encontrar el camino que minimice $D_p(X,Y)$, siendo una elección natural el tomar el mínimo sobre todos los posibles caminos:

$$D(X,Y) = min_P(D_P(X,Y)) \tag{3.96}$$

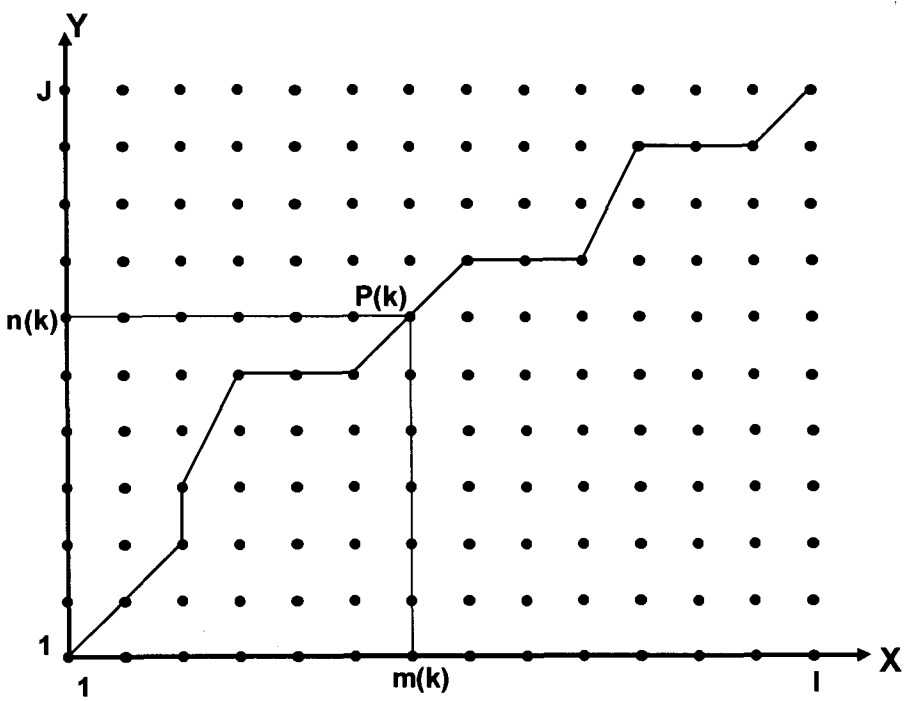

Figura 3.9. Ejemplo de camino de distorsión temporal para los patrones de voz X e Y.

Para resolver este problema se emplean técnicas inspiradas en los algoritmos de programación dinámica. Con objeto de tener en cuenta ciertos aspectos físicos del problema y limitar el número de caminos a considerar, se imponen algunas restricciones a las funciones de distorsión:

- Puntos de inicio y final de tramas.

$$P(1) = (1,1) \tag{3.97}$$

$$P(K) = (I, J) \tag{3.98}$$

- El camino no podrá tener una pendiente negativa.
- Continuidad Local. Para minimizar la pérdida de información se restringen los movimientos locales. Un ejemplo de restricciones recogidas en [RAB93] es:

$$g(i,j) = min\begin{cases} g(i-1,j) + d(i,j) \\ g(i-1,j-1) + 2d(i,j) \\ g(i,j-1) + d(i,j) \end{cases} \tag{3.99}$$

- Restricciones de pendiente. Se expresan como funciones de costo para el cálculo del peso w(k). Algunos ejemplos que aparecen en [SAK78] son:

$$\mathrm{w(k)} = n(k) - n(k-1) + m(k) - m(k-1) \tag{3.100}$$

$$\mathrm{w(k)} = n(k) - n(k-1) \tag{3.101}$$

$$\mathrm{w(k)} = m(k) - m(k-1) \tag{3.102}$$

$$\mathrm{w(k)} = \{\, n(k) - n(k-1) + m(k) - m(k-1) \,\} \tag{3.103}$$

Por su parte, el factor de normalización suele ser:

$$N(w) = I + J \tag{3.104}$$

Finalmente, para el caso descrito anteriormente, el algoritmo de programación dinámica puede expresarse de la siguiente manera:

1. Inicio:

$$g(1,1) = d(1,1) \cdot w(1)$$

2. Recursión:

Para $1 \leq i \leq I$

Para $1 \leq j \leq J$

$$g(i,j) = min\begin{cases} g(i-1,j) + d(i,j) \\ g(i-1,j-1) + 2d(i,j) \\ g(i,j-1) + d(i,j) \end{cases}$$

3. Terminación:

$$D(X,Y) = g(I,J) / (I+J)$$

Las técnicas de distorsión temporal se han aplicado con éxito a los problemas de reconocimiento de palabras aisladas y conectadas (series de palabras concatenadas sin la presencia de pausas artificiales entre ellas). Sin embargo para este último caso, la potencia computacional requerida puede ser excesiva. Con el objetivo de aliviar este problema se han propuesto diferentes técnicas:

- Algoritmo de programación dinámica en dos niveles (*two-level dynamic programming algorithm*). La idea es dividir el proceso de encaje de patrones en dos fases. Durante la primera se trabaja en el nivel de palabra y durante la segunda se trata el nivel de habla conectada.

- Algoritmo de construcción de nivel (*level building algorithm*). Es una implementación más eficaz que la anterior, conseguida a través de la realización de decisiones parciales sobre las palabras durante la programación dinámica.

- Algoritmo de pasada única (*one-pass algorithm*). Resulta muy útil para aplicaciones en tiempo real ya que su computación se realiza trama a trama de forma síncrona.

3.4.2. Redes neuronales

Una red neuronal [LIP87a] [KOH90] [ROB94] [TEB95] es una estructura de cómputo capaz de discriminar y modelar características no lineales. Está compuesta por un conjunto, usualmente grande, de unidades de procesamiento sencillas conectadas entre sí por una serie de conexiones. Todas estas unidades operan de forma simultánea, constituyendo un esquema de paralelismo masivo de grano fino. Aparte de las unidades de proceso o neuronas no existe otro tipo de unidad de control o procesamiento adicional.

Las neuronas de una red se dividen en unidades de entrada, unidades ocultas y unidades de salida. Las células de entrada y salida, como su nombre indica, reciben los datos de entrada al sistema y los resultados proporcionadas por la red, respectivamente. Por su parte, las neuronas de la capa oculta son las encargadas de la transformación de las diferentes representaciones internas de los datos.

El tipo de conexión entre unidades de proceso determina en buena parte el uso de la red neuronal. Aunque una red puede presentar cualquier tipo de topología las más usuales son (Figura 3.10):

1. Redes desestructuradas. Son útiles para tareas de completado de patrones.
2. Redes organizadas en niveles. Empleadas para asociación de patrones.
3. Redes recurrentes. Empleadas para secuenciación de patrones a lo largo del tiempo.
4. Redes modulares. Permiten construir sistemas complejos a partir de componentes más simples.

El funcionamiento de la red viene determinado por el modo de computación de las unidades de proceso. La entrada de una célula está formada por la combinación de los valores de salida de las neuronas que la preceden, a través de la fórmula:

$$x_j = \sum_i w_{ij} y_j + \theta_j \tag{3.105}$$

Los valores w_{ij} son los pesos de interconexión entre células de la red y el término θ_j representa un valor constante para la conexión j. En la práctica, este término se considera como una conexión extra donde el peso es precisamente ese valor y la entrada de la conexión vale 1.

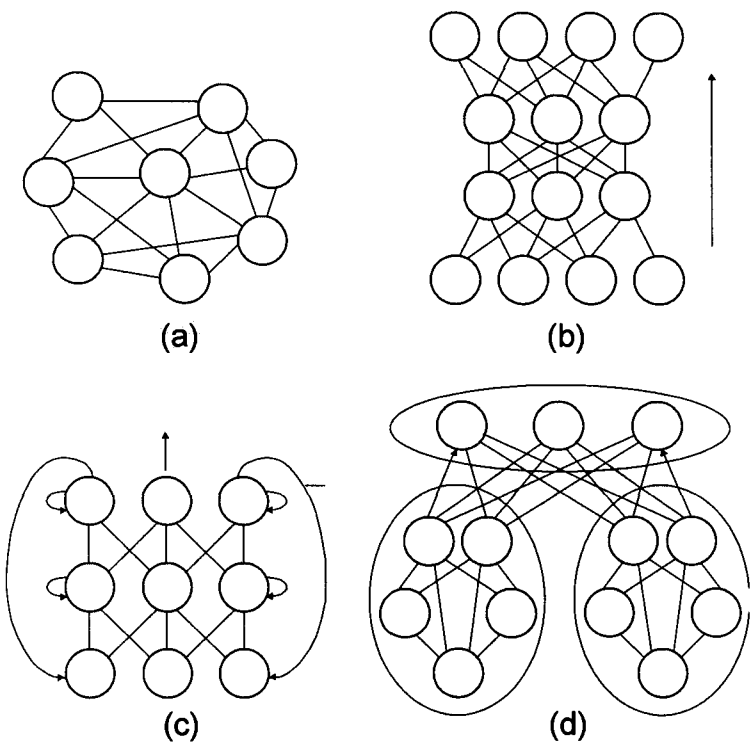

Figura 3.10. Topologías de redes neuronales: (a) Desestructurada. (b) Organizadas en niveles. (c) Recurrente. (d) Modular.

La célula aplica a la entrada anterior una función que puede ser de tipo lineal o no lineal. A su vez dentro del grupo de las no lineales encontramos funciones umbral y funciones sigmoide. El caso lineal no se usa mucho porque proporciona muy poca potencia de cálculo. Un ejemplo de función umbral sencillo lo constituye la función escalón.

Las funciones sigmoide suelen ser las más empleadas, siendo algunos ejemplos:

$$y = \frac{1}{1+\exp(-x)} \tag{3.106}$$

$$y = \tanh(x) \tag{3.107}$$

El último factor que define a una red neuronal es el tipo de entrenamiento que se realiza. El proceso de entrenamiento tiene por objeto modificar el valor de los pesos de interconexión de la red reforzando aquellas conexiones entre dos unidades con salidas correladas. Este mecanismo se conoce como regla de Hebb (donde además se define un factor de aprendizaje ε):

$$\Delta w_{ij} = \varepsilon y_i y_j \tag{3.108}$$

En cada caso se emplearán diferentes variantes de la anterior regla, siendo una de las más conocidas el método de retropropagación del error (*Backpropagation*).

Dependiendo del modo en que se entrena una red neuronal se pueden clasificar en:

1. Redes con aprendizaje supervisado. Al mismo tiempo que se proporcionan las entradas al sistema se incluye la salida deseada que la red debe producir.
2. Redes con aprendizaje semi-supervisado. Junto a la entrada, solamente se indica si el patrón se ha clasificado bien o mal durante el proceso de entrenamiento.
3. Redes con aprendizaje no supervisado. La red debe encontrar por sí misma la regularidad presente en los datos de entrada y ordenarlos en consecuencia.

Aplicadas a problemas de reconocimiento de voz [WAI88], se utilizan muy diversos esquemas dedicados en la mayoría de los casos a la detección de fonemas [WAI89]. Destaca también la aplicación de modelos de red neuronal en conjunción con otros mecanismos como son los modelos ocultos de Markov, dando lugar a sistemas híbridos [MOR95].

3.4.3. Modelos ocultos de Markov

Un modelo oculto de Markov o HMM (*Hidden Markov Model*) [BAK75], [BAH83], [RAB86], [POR88], [RAB89], [PIC90], es una colección de estados conectado por transiciones (Figura 3.11).

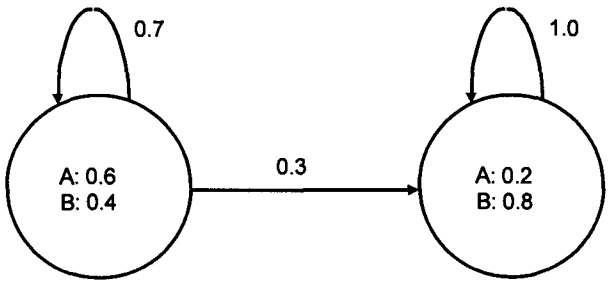

Figura 3.11. Modelo oculto de Markov con 2 estados y 2 símbolos de salida A y B.

El proceso comienza en un estado diseñado para tal efecto y en cada paso de ejecución se toma una transición a un nuevo estado y se genera un símbolo de salida asociado a ese estado. La elección de la transición y del símbolo es aleatoria y está gobernada por distribuciones de probabilidad. Un modelo oculto de Markov puede considerarse como un caja negra donde la secuencia de símbolos de salida generados a lo largo del tiempo es visible, pero la secuencia de estados por los que se ha pasado para generar la anterior serie se desconoce. Esto es por lo que se llaman modelos ocultos.

Cuando se aplican los modelos ocultos de Markov al reconocimiento de la voz, los estados se interpretan como modelos acústicos, indicando las ocurrencias de sonidos que son más probables durante los correspondientes segmentos de habla. Por su parte las transiciones incorporan restricciones de tipo temporal acerca de cómo son las secuencias de aparición de esos sonidos. Estados y transiciones pueden ser utilizados para modelar distintas jerarquías del proceso del habla: desde fonemas hasta oraciones pasando por palabras.

La utilización de este tipo de estructuras debe resolver 3 problemas básicos:

1).- Problema de reconocimiento. Consiste en escoger aquel modelo de entre un grupo de éstos, que mejor represente al conjunto de etiquetas obtenidas a partir de la cuantificación de las correspondientes plantillas espectrales.

Formalmente, conocida una secuencia de observaciones $O= \{o_1, o_2, \ldots, o_T\}$ y el correspondiente modelo de Markov $\lambda = (A, B, \pi)$, siendo $A= \{a_{ij}\}$ la distribución de probabilidad para la transición entre los estados i y j, $B= \{b_j(k)\}$ la distribución de probabilidad para la observación k en el estado j y π la distribución inicial de estados, se pretende calcular $Prob(O \mid \lambda)$.

Una solución eficiente a este problema es el algoritmo hacia adelante y atrás (*forward-backward algorithm*) [BAU70], [RAB86]:

$$\text{Prob}(O \mid \lambda) = \sum_{i=1}^{N} \alpha_T(i) \tag{3.109}$$

$$\alpha_1(i) = \chi_i b_i(o_1), \quad 1 \le i \le N \tag{3.110}$$

$$\alpha_{t+1}(j) = \left[\sum_{i=1}^{N} \alpha_t(i) a_{ij} \right] b_j(o_{t+1}), \quad \begin{matrix} 1 \le t \le T-1 \\ 1 \le j \le N \end{matrix} \tag{3.111}$$

2).- Problema de decodificación. Descubrir la secuencia oculta de estados, es decir, conocida una secuencia de observaciones $O = \{ o_1, o_2, \ldots, o_T \}$ calcular la secuencia de estados $I = \{ i_1, i_2, \ldots, i_T \}$ que sea óptima siguiendo algún criterio.

Este segundo problema puede resolverse a partir del algoritmo de Viterbi [VIT67]. Este algoritmo es similar al anterior con la excepción de que en vez de tomar la suma de los valores de probabilidad en los anteriores estados se toma el máximo. Para ello se define una función de máxima probabilidad $\delta_t(i)$, siendo i el estado y t el instante de tiempo:

$$\delta_1(i) = \chi_i b_i(o_1) \tag{3.112}$$

$$\delta_t(j)=\left[\max_{1\le i\le N}\left(\delta_{t-1}(i)\cdot a_{ij}\right)\right]\cdot b_j(o_t) \quad \begin{matrix} 2\le t\le T \\ 1\le j\le N \end{matrix} \tag{3.113}$$

3).- Problema de aprendizaje o entrenamiento. Consiste en construir un modelo de manera que recoja el conocimiento con el que se ha entrenado de forma óptima, es decir, construir el modelo $\lambda = (A, B, \pi)$ para que $Prob(O \mid \lambda)$ sea máxima.

El caso se resuelve a través de un algoritmo iterativo de tipo gradiente que se conoce como algoritmo Baum-Welch [BAU70], [BAH83]. Este procedimiento proporciona un método para reestimar unos nuevos valores de a'_{ij} y $b'_j(k)$ de forma que se cumpla que $Prob(O \mid \lambda') \ge Prob(O \mid \lambda)$:

$$a'_{ij}=\frac{\sum_{t=1}^{T}\gamma_t(i,j)}{\sum_{t=1}^{T}\sum_{j=1}^{N}\gamma_t(i,j)} \tag{3.114}$$

$$b'_j(k)=\frac{\sum_{t=1}^{T}\sum_{j=1}^{N}\gamma_t(i,j)}{\sum_{t=1}^{T}\sum_{j=1}^{N}\gamma_t(i,j)} \quad o_t=k \tag{3.115}$$

$$\gamma_t(i,j)=\frac{\alpha_t(i)\cdot a_{ij}b_j(o_{t+1})\cdot\beta_{t+1}(j)}{P(O|\lambda)} \tag{3.116}$$

$$\beta_t(i)=\sum_{j=1}^{N}a_{ij}b_j(o_{t+1})\cdot\beta_{t+1}(j) \tag{3.117}$$

El esquema de modelos ocultos de Markov que hemos considerado hasta ahora corresponde a una situación en la cual el espacio acústico se ha dividido en un número moderado de regiones como resultado del proceso de cuantificación vectorial. Así la distribución de probabilidad a lo largo del espacio acústico se representa únicamente por un simple histograma para cada una de las entradas del libro de códigos empleado. Este enfoque, que se conoce como de modelos de densidad discreta, es computacionalmente muy efectivo pero presenta el problema de decidir el tamaño de los libros de códigos.

Un tamaño pequeño introducirá mucho ruido de cuantificación y un tamaño grande puede hacer que muchos centroides no estén lo suficientemente representados en los datos de entrenamiento y producir con ello una degradación en las tasas de reconocimiento.

Una solución a este problema puede ser el empleo de modelos de densidad continua [LIP82]. En este caso la distribución de probabilidad a través del espacio acústico se modela directamente asumiendo que presenta una determinada forma paramétrica, generalmente en la forma de una mezcla de K distribuciones Gaussianas [RAB89].

$$b_j(y) = \sum_{k=1}^{K} c_{jk} G\left(y, \mu_{jk}, C_{jk}\right) \tag{3.118}$$

En este caso los parámetros $b_j(y)$ dependen de un conjunto de pesos c_{jk}, que ponderan a cada Gaussiana G de media μ_{jk} y covarianza C_{jk}. La suma de todos los pesos de ponderación de las Gaussianas debe ser igual a 1.

Este tipo de modelos proporciona para problemas con un gran número de vocabulario el mejor resultado, aunque con un coste computacional de 2 órdenes de magnitud mayor. El principal inconveniente que presenta esta aproximación es que los parámetros no están compartidos por todos los estados, de tal forma que si el valor de K es grande podemos tener una falta de representatividad para muchos de los estados y un valor pequeño de K haría errónea la asunción de que la distribución de probabilidad puede modelarse por medio de una mezcla de Gaussianas.

Por último existe un caso intermedio que se conoce como modelos de densidad semicontinua [HUA90]. En este caso existe un libro de códigos común para todos los estados, que describe el conjunto de clases acústicas. Sin embargo aquí las clases no se representan por centroides discretos sino por medio de funciones de densidad continua (típicamente Gaussianas) a través del espacio de una clase. De esta forma se evitan los errores de cuantificación que se producen cuando un vector se transforma automáticamente en el centroide de la clase a la que pertenece. Al mismo tiempo este enfoque proporciona una solución mejor en cuanto a tiempos de ejecución.

$$b_j(y) = \sum_{k=1}^{L} c_{jk} G(y, \mu_k, C_k) \tag{3.119}$$

El número de elementos del libro de códigos aparece representado por L y se corresponde también con el número de Gaussianas. Estas funciones como en el caso anterior se recalculan durante el entrenamiento al mismo tiempo que los parámetros de los modelos ocultos. Esta es otra diferencia con los modelos discretos donde el libro de códigos permanece fijo durante todo el proceso.

Para los diferentes tipos de estructuras de los HMM estudiadas, encontramos dos variantes que merece la pena comentar:

La primera variante consiste en asociar las observaciones con los arcos mas que con los estados de un HMM [BAH83]. En este tipo de modelo, podemos definir transiciones nulas que no producen ninguna salida. Estas transiciones proporcionan una manera eficaz de describir fenómenos de elisión fonética.

La segunda variante se corresponde con el concepto de unión de parámetros (*parameter tying*) [BAH83]. La idea es reducir el número de parámetros independientes por medio del establecimiento de relaciones de equivalencia entre parámetros pertenecientes a diferentes estados. Este método es particularmente interesante cuando la cantidad de datos de entrenamiento para generar los modelos es insuficiente. Es importante mencionar que los aspectos matemáticos del algoritmo de entrenamiento no se ven afectados por la ligadura de parámetros.

Otro aspecto importante en los trabajos de investigación relacionados con los HMM es el punto relativo al modelado del tiempo de permanencia en un estado. Lo usual es obtener las distribuciones de probabilidad que modelan la duración por cada estado a partir de los datos de entrenamiento. Estas probabilidades se emplean durante una fase de postproceso con objeto de matizar las puntuaciones obtenidas en el algoritmo de Viterbi. Esto puede ser especialmente útil en habla continua [ANA95], [BUR96].

Aunque el marco teórico para incorporar informaciones referentes al tiempo

dentro de los HMM está bien desarrollado [RAB89], su coste computacional es usualmente demasiado alto.

El entrenamiento de los modelos ocultos de Markov, tal y como lo hemos desarrollado, se realiza siguiendo un criterio de máxima semejanza o ML (*Maximum Likehood*), es decir, se intenta maximizar la probabilidad de una secuencia(s) de observaciones siendo conocido el modelo asociado. Aunque este criterio posee buenas propiedades asintóticas, requiere por lo general de grandes conjuntos de entrenamiento para alcanzar una estimación fiable. Para resolver posibles problemas asociados con la escasez de datos, encontramos técnicas de suavizado tales como la interpolación borrada (*deleted interpolation*) [JEL80].

Sin embargo, encontramos en la literatura otros criterios alternativos:

1. **Criterio de máxima información mutua o MMI** (*Maximum Mutual Information*) [BAH86]. Se basa en la idea de diseñar todos los HMM al mismo tiempo, de forma que las capacidades discriminativas de cada modelo puedan ser aumentadas. Suponiendo que las unidades de habla representadas por medio de los modelos HMM son equiprobables, este criterio estima los parámetros HMM haciendo máximo:

$$I_i = max_\lambda \left(\log P\left(O^i|\lambda_i\right) - \log \sum_{j=1}^{M} P\left(O^i|\lambda_j\right) \right) \tag{3.120}$$

siendo M el número de modelos, λ_i el modelo correcto y λ_j el resto de modelos. Cuando se aplica a todas las secuencias de entrenamiento, la expresión queda de la siguiente forma:

$$I = max_\lambda \left(\sum_{i=1}^{M} \left(\log P\left(O^i|\lambda_i\right) - \log \sum_{j=1}^{M} P\left(O^i|\lambda_j\right) \right) \right) \tag{3.121}$$

El algoritmo de entrenamiento Baum-Welch puede también extenderse al caso MMI [GOP89], [NOR91]. Así mismo, este tipo de enfoque puede utilizarse en sistemas híbridos red neuronal/HMM [RIG94].

2. **Criterio de mínima información discriminante o MDI** (*Maximum*

Discrimination Information) [EPH89]. En este caso, la mejora en el proceso de entrenamiento de los HMM se realiza mediante la selección de los parámetros que minimizan la información de discriminación entre la distribución de probabilidad de la fuente (conjunto de observaciones) y la del modelo de oculto de Markov. La información de discriminación entre dos distribuciones de probabilidad P y Q, con funciones de densidad de probabilidad p y q, puede expresarse por medio de la siguiente ecuación: [EPH87]:

$$D(Q||P) = \int q(y) \ln\left(\frac{q(y)}{p(y)}\right) dy \tag{3.122}$$

La idea detrás del criterio MDI es compensar los desajustes entre las medidas y el modelo. El procedimiento comienza con la estimación de los HMM siguiendo el criterio ML. Entonces para un HMM dado se estima la distribución de probabilidad de la fuente, minimizando la información de discriminación sobre todas las distribuciones de probabilidad de la fuente que son coherentes con las medidas. Finalmente, dada una distribución de probabilidad de la fuente, se estima el HMM que minimiza la información de discriminación sobre el conjunto de modelos ocultos de Markov. La reestimación se realiza usando una variante modificada del algoritmo Baum-Welch [EPH87].

3. **Criterio del mínimo error en la clasificación o MEC** (Minimum Error Classification) [BAH88a], [JUA92], [REI95], [RAH97]. Este método conocido también como entrenamiento correctivo (*corrective training*) o entrenamiento discriminativo (*discriminative training*), constituye una fase de postproceso que busca aumentar el poder de discriminación de los modelos por medio de la reestimación de los parámetros de éstos. El procedimiento se centra en las porciones de los HMM, que más importancia tienen para la discriminación entre unidades similares que deben ser reconocidas. Consiste en un paso adicional comparado con el ML, puesto que el procedimiento de entrenamiento no necesita ser reformulado.

Este modelo también puede aplicarse a sistemas que empleen redes de neuronas [LEE95].

El modelado de los modelos ocultos de Markov de primer orden, asume que la probabilidad de transición entre estados en el momento *t+1* depende solamente del estado de la cadena de Markov en el tiempo *t*. En los modelos de segundo orden HMM2 [MAR94], [MAR97] la probabilidad de transición de estado en el instante *t+1* depende de los estados de la cadena en los momentos *t* y *t-1*. Los experimentos llevados a cabo en tareas de reconocimiento de dígitos conectados muestran una mejora en las prestaciones en el caso de los HMM2. Sin embargo si al caso HMM (HMM1) se le añade una etapa de postproceso que tenga en cuenta la duración de los segmentos [GON94], los resultados son similares.

Los modelos ocultos de Markov, tanto de primer como de segundo orden, sufren de una limitación conocida como la asunción de la independencia de las observaciones. Con ello se asume que la probabilidad de un vector de características acústicas, asociado a un estado particular, no depende de los otros vectores de ese mismo estado. En la mayoría de los casos esta asunción de independencia no es válida.

El uso de propiedades dinámicas a corto plazo en los vectores de rasgos ayuda a solventar este problema. Sin embargo, diversos estudios apuntan directamente hacia el mismo. La dependencia estadística entre el vector de observación actual y el último observado puede modelarse explícitamente por medio estimación gausiana [WEL87]. Otra posibilidad es incluir un método que condicione la probabilidad del vector de observación actual en el estado actual con los vectores de observación anteriores [PAL93]. Las restricciones secuenciales también se pueden capturar por medio de una red neuronal recurrente [ROB94].

Una traza de voz se representa como un punto que se mueve en el espacio de parámetros según se producen cambios articulatorios. Esto constituye la trayectoria del habla. Teniendo en cuenta que un punto puede pertenecer a diferentes trayectorias, los modelos para reconocimiento de voz deben basarse en las trayectorias y no en las posiciones geométricas individuales del espacio de parámetros. La independencia de observaciones presente en los HMM no preserva la información de trayectorias. Más exactamente, diferentes trayectorias comparten la misma función de distribución de probabilidad. Las clases con las trayectorias no pueden representarse bien, porque la

información sobre la continuidad de cada trayectoria individual se pierde. Las trayectorias se agrupan y esto conlleva a una menor discriminación de contextos fonéticos complejos.

Una forma de tener en cuenta este aspecto es modelar la función de densidad de probabilidad de un gran vector, compuesto por la concatencación de las tramas de la secuencia de los vectores de observación, a través de una distribución gausiana multivariada. En otras ocasiones, las mezclas se introducen en el nivel de segmentos. De esta forma, una distribución gausiana representa una clase simple de trayectorias, mientras que un modelo de mezcla puede representar múltiples clases de agrupación de trayectorias.

Una nueva aproximación consiste en considerar a las trayectorias como observaciones de una variable aleatoria y emplear un modelo de trayectorias con mezcla estocástica o STM (*Stochastic Mixture Trajectory*) [GON94], [OST96]. De esta forma, las unidades del habla basadas en fonemas, se modelan como agrupaciones de trayectorias en su espacio de parámetros. A su vez, las trayectorias se modelan por medio de mezclas de secuencias de estados de las funciones de densidad Gaussianas multivariadas y se optimizan a nivel de secuencias de estados. Por último, la duración de las trayectorias se estima por medio de distribuciones Γ.

3.5. MODELOS DE LENGUA

La tarea que debe realizar un sistema automático de reconocimiento de voz es encontrar la cadena de palabras que satisfaga:

$$\hat{W} = \arg\max_{W} P(A|W)P(W) \tag{3.123}$$

donde A son los datos acústicos y $W = w_1, w_1, \ldots, w_n$, con $w_i \in V$, denota la cadena de n palabras de entre un vocabulario de tamaño fijo V.

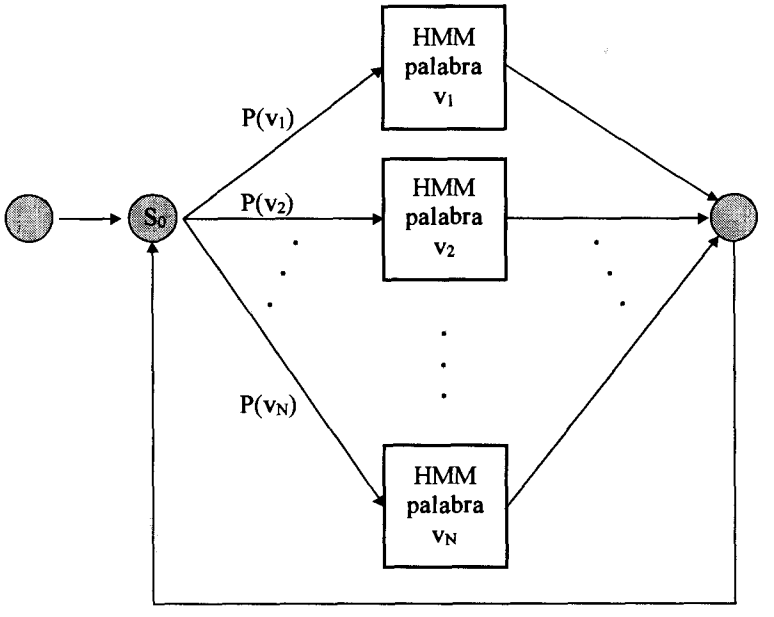

Figura 3.12. Modelo compuesto de producción de discurso cuando la generación de palabras no sigue ninguna gramática.

Un modelo de lengua es el mecanismo que permite asignar a cada posible secuencia de palabras W, la probabilidad asociada $P(W)$. En este caso una palabra queda definida por su pronunciación. De esta forma, si una palabra tiene varias pronunciaciones se considerará que son dos entidades diferentes. De igual manera los vocablos homófonos se considerarán una única palabra.

Si empleamos la regla de Bayes, el valor de P(W) puede descomponerse de manera formal en:

$$P(W) = \prod_{i=1}^{n} P(w_i | w_1, \ldots, w_{i-1}) \tag{3.124}$$

donde $P(w_i|w_1, \ldots, w_{i-1})$ es la probabilidad de que w_i sea realizada teniendo en cuenta que las palabras $w_1, \ldots, w_{i-1}$ fueron pronunciadas previamente. Estas n últimas palabras se conocen como historia y suelen denotarse por h_i.

En la práctica sin embargo, no es posible estimar los valores $P(w_i|w_1, \ldots, w_{i-1})$. Suponiendo que tengamos un vocabulario de tamaño $|V|$, podremos tener $|V|^{i-1}$ historias diferentes, lo cual para vocabularios de varios miles de palabras es inabordable. La solución a este problema pasa por agrupar todas las posibles historias en un número manejable de clases de equivalencia [JEL92]. La idea que sigue este planteamiento es que incluso para valores pequeños de i, muchas combinaciones de esas i palabras no se producen en la realidad.

Definimos la función Φ como la aplicación de un conjunto de historias en alguna de las clases de equivalencia anteriormente citadas. Si $\Phi(w_1, w_2, w_{i-1})$ denota la clase de equivalencia de la cadena w_1, w_2, w_{i-1} entonces la probabilidad $P(W)$ puede aproximarse por medio de la siguiente expresión:

$$P(W) = \prod_{i=1}^{n} P(w_i | \Phi(w_1, \ldots, w_{i-1})) \tag{3.125}$$

Viendo el conjunto de clases como un autómata de estados finitos, podemos considerar que la introducción de una nueva palabra hace evolucionar a la gramática del estado Φ_{t-1}, (instante $i-1$) al estado Φ_t. De esta forma la ecuación anterior puede expresarse como:

$$P(W) = \prod_{i=1}^{n} P(w_i | \Phi_{i-1}) \tag{3.126}$$

Con objeto de poder evaluar los términos $P(w_i| \Phi_{t-1})$, definimos $C(w, \Phi)$ como el número de veces que la palabra w alimenta al autómata de estados inmediatamente después de que éste se encuentre en el estado Φ. De igual manera $C(\Phi)$ denota el número de veces que el autómata alcanza el estado Φ:

$$C(\Phi) = \sum_{w} C(w, \Phi) \tag{3.127}$$

De esta forma la estimación de la probabilidad buscada queda:

$$P(w_i | \Phi_i = \Phi) = \frac{C(w, \Phi)}{C(\Phi)} \tag{3.128}$$

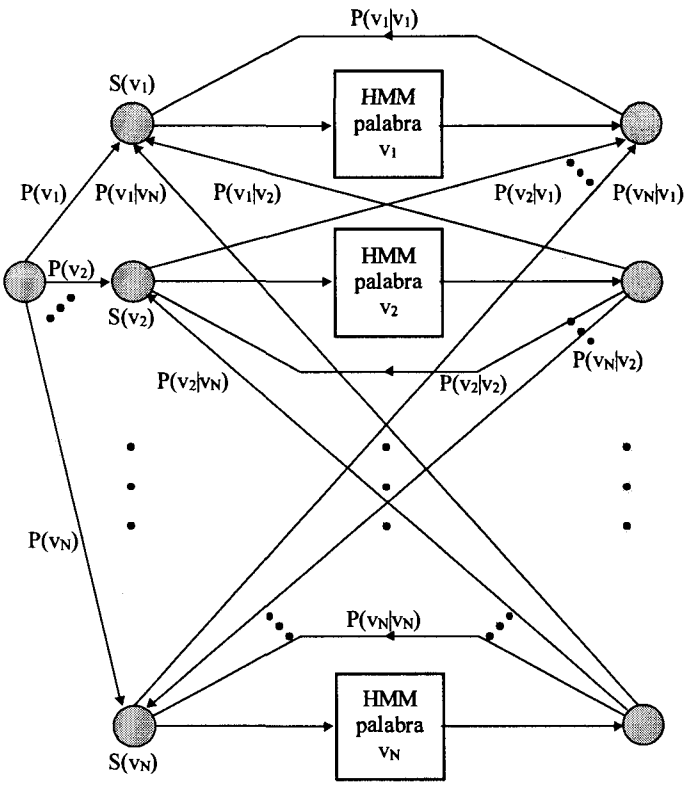

Figura 3.13. Modelo compuesto de producción de discurso cuando la generación de palabras depende solamente de la identidad de la palabra precedente (modelo de lengua de bigramas).

En la práctica los valores *C(w, Φ)* se obtienen a partir de grandes bases de datos con documentos de texto. Usualmente el criterio para determinar las clases de equivalencia suele ser lo que se conoce como N-gramas (*N-grams*), es decir, dos historias son equivalentes si las *N-1* palabras coinciden. Los casos más comunes son N=2 o bigramas (*bigrams*) [NEY92a] y N=3 o trigramas (*trigrams*) [JEL85]. La fórmula para el caso de los trigramas es por tanto:

$$P(W) = \prod_{i=1}^{n} P(w_i | w_{i-2}, w_{i-1}) \tag{3.129}$$

$$P(w_3 | w_1, w_2) = f(w_3 | w_1, w_2) \doteq \frac{C(w_1, w_1, w_3)}{C(w_1, w_2)} \tag{3.130}$$

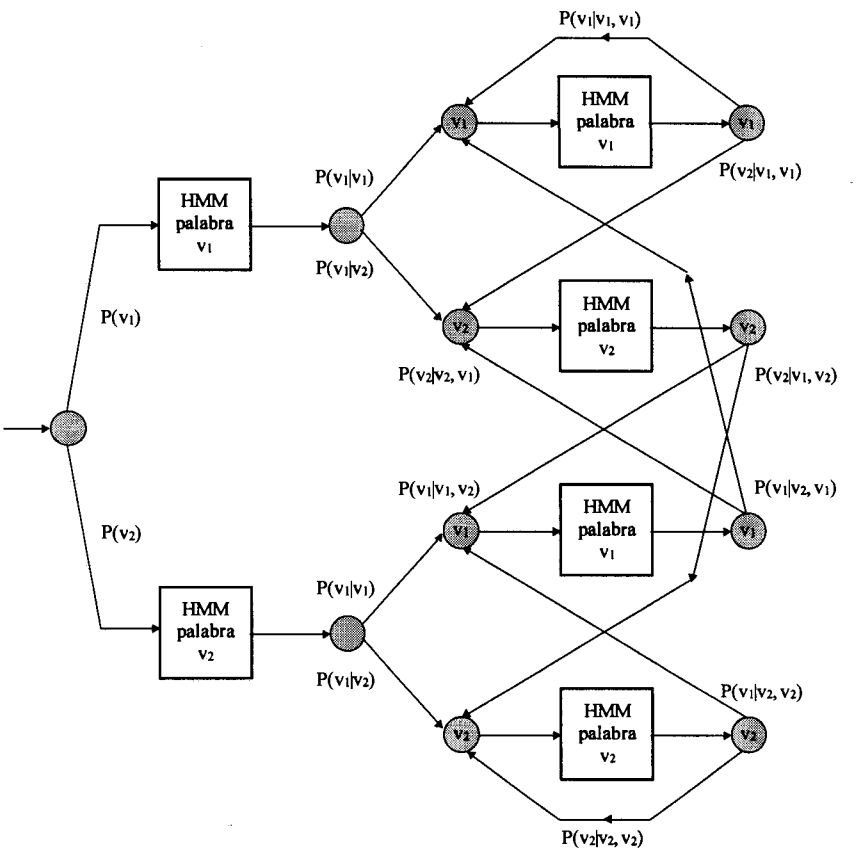

Figura 3.14. Modelo compuesto de producción de discurso basado en un modelo de lengua con trigramas cuando el vocabulario consta de dos palabras.

Uno de los problemas con los trigramas es que para sistemas con vocabularios muy grandes (200.000 palabras), la red resultante puede no resultar factible. Una solución a este problema aparece en [DER86]. En este caso se emplea como modelo de lengua uno que denominan tri-POS (*Parts Of Speech*) y un conjunto de restricciones sintácticas globales empleando un *parser* de oraciones. La gran ventaja de este método es la importante disminución en el espacio de memoria requerido para almacenar la red fonética, así como, la necesidad de una menor cantidad de datos de entrenamiento. Sin embargo, el cálculo de la probabilidad condicionada a partes de la red de las 2 palabras

anteriores proporciona unas restricciones lingüísticas mucho más débiles que las del modelo de trigramas.

Sin embargo y debido a la falta de suficientes datos (documentos de texto), muchas de las combinaciones de N-palabras no aparecerán en el conjunto de textos disponibles con lo que la probabilidad asociada será 0. Se hace necesario algún método que permita suavizar la estimación de probabilidad para eventos no presentes [ESS92]. Una forma de hacerlo es mediante la interpolación lineal de las frecuencias de trigramas, bigramas y unigramas (*deleted interpolation*) [JEL92]:

$$P(w_3|w_1,w_2) = \lambda_3 f(w_3|w_1,w_2) + \lambda_2 f(w_3|w_2) + \lambda_1 f(w_3) \tag{3.131}$$

$$\lambda_1 + \lambda_2 + \lambda_3 = 1 \tag{3.132}$$

Los pesos de la interpolación lineal λ_i se estiman haciendo máxima la probabilidad para un conjunto de datos (textos) diferente de los empleados para calcular la frecuencia de aparición de los n-gramas. El algoritmo de Baum-Welch puede emplearse perfectamente para hallar la solución de este problema de máxima semejanza.

Otra técnica que sirve para suplir la falta de datos de entrenamiento es el método de retroceso (*backing-off*) [KAT87]. Este método es el que prevalece en muchos reconocedores actuales.

$$\hat{P}(w_3|w_1,w_2) = \begin{cases} f(w_3|w_1,w_2)w_2 & si\ C(w_2,w_3) \geq K \\ \alpha Q_T(w_3|w_1,w_2) & si\ 1 \leq C(w_2,w_3) \geq K \\ \beta(w_1,w_w)\hat{P}(w_3,w_2) & resto\ de\ casos \end{cases} \tag{3.133}$$

Los valores α y β se escogen, de forma que la probabilidad quede normalizada de manera adecuada. $Q_t(w_3|\ w_1,\ w_2)$ es una función del tipo Good-Turing [GOO53] y la estimación de probabilidad de $P(w_3|\ w_2)$ es la estimación de probabilidad del bigrama que se realiza de la misma forma que la estimación de $P(w_3|\ w_1,\ w_2)$:

$$\hat{P}(w_3|w_2)=\begin{cases} f(w_3|w_2) & si\ C(w_2,w_3)\geq L \\ \alpha Q_T(w_3|w_2) & si\ 1\leq C(w_2,w_3)\geq L \\ \beta(w_2)f(w_3) & resto\ de\ casos \end{cases} \tag{3.134}$$

siendo los valores de L y M umbrales determinados de forma intuitiva.

La idea detrás de este método, es que si hay suficiente número de casos, la frecuencia relativa es una buena estimación de la probabilidad. Si no, se debe retroceder y calcular estas probabilidades tomando las frecuencias de los bigramas y llegado el caso a partir de los unigramas.

En general no resulta práctico emplear n-gramas para valores de n mayores de 3. La ligera ganancia en la complejidad de decisión de la gramática, se ve claramente contrarrestada por la mayor dificultad que presenta la estructura de decisión. Algunos trabajos además intentan añadir criterios sintácticos, como la categoría que ocupan las palabras en la oración (sustantivos, verbos, adjetivos, etc.), con el fin de disminuir la complejidad de decisión y con el propósito de mejorar los procedimientos de suavizado [MAL92]. Incluso puede permitirse la existencia de palabras que aparezcan en diversas categorías, como consecuencia de las distintas funciones gramaticales o significados radicalmente diferentes que pueden presentar éstas [JAR96].

Algunas mejoras en el modelado tienen en cuenta la historia de ocurrencias en las palabras, modelando la mayor o menor concentración en determinados textos o partes de un texto de manera dinámica [LAU93]. En otros casos, en vez de emplear la identidad de las palabras más recientes para definir las clases de equivalencia de una historia, se usa el estado de un *parser* gramatical con objeto de definir el evento condicionante [GOD92].

Por otra parte, si además queremos disponer de conocimiento semántico, el proceso requerirá etapas adicionales para aprender y extraer de forma automática dichas estructuras semánticas [KUH94].

3.6. BÚSQUEDA DE HIPÓTESIS

Debemos buscar ahora todas las posibles cadenas de palabras W para encontrar la que maximiza la fórmula:

$$\hat{W} = \arg\max_{W} P(A|W)P(W) \tag{3.135}$$

donde A son los datos acústicos y $W= w_1, w_1, ..., w_n$, con $w_i \in V$, denota la cadena de n palabras de entre un vocabulario de tamaño fijo V.

La búsqueda no puede ser conducida por la fuerza bruta, puesto que el espacio de W es enormemente grande. Esto quiere decir que se necesita un conjunto reducido de hipótesis de búsqueda que no examinará el contundente número de posibles candidatos y solamente tendrá en cuenta aquellas cadenas de palabras que de alguna manera estén sugeridas por la estructura acústica A.

La tarea de encontrar la oración que mejor satisface todas las restricciones acústicas y lingüísticas puede realizarse mediante el uso de dos estrategias básicas: integrada y modular.

En la aproximación integrada, la decisión implicada en el proceso de reconocer se realiza considerando todas las fuentes de conocimiento de manera combinada. En principio, esta estrategia consigue los mejores resultados si todas las fuentes de conocimiento pueden caracterizarse e integrarse de manera completa. En este caso es posible aproximarse a la jerarquía de conocimiento lingüístico (acústico, léxico, sintáctico y semántico) y compilarlas dentro de una red de estados finitos compuesta por nodos gramaticales, nodos correspondientes a los modelos ocultos de Markov y las conexiones entre ellos.

De esta forma, el problema del reconocimiento se resuelve mediante el encaje de la secuencia de vectores de características de entrada con la mejor secuencia de palabras que atraviese la citada red de conocimiento. Esta es la estrategia de búsqueda que normalmente adoptan los sistemas de reconocimiento de hoy en día. Sin embargo,

encontramos una serie de problemas con este tipo de aproximación. El primero es que no todas las fuentes de conocimiento pueden caracterizarse e integrarse de una forma total; un buen ejemplo puede ser la prosodia. Aparte para muchas tareas donde el tamaño del vocabulario es grande la red resultante puede ser computacionalmente intratable.

La aproximación modular mostrada en la (Figura 3.15) emplea la información proporcionada por las diferentes fuentes de conocimiento de una forma secuencial, pudiéndose especificar cada módulo de forma separada. La mayoría de los sistemas de diálogo funcionan con este principio de colaboración y su mayor ventaja radica en que es computacionalmente una opción más factible. Por contra, la mayor limitación de este modelo es que las decisiones que se realizan en una etapa no tienen en cuenta a las otras etapas/fuentes de conocimiento. Así, los errores de decisión se propagan de una etapa a la siguiente y esta acumulación es susceptible de causar errores en la búsqueda a menos que, se mantengan para cada paso del proceso de decodificación múltiples hipótesis.

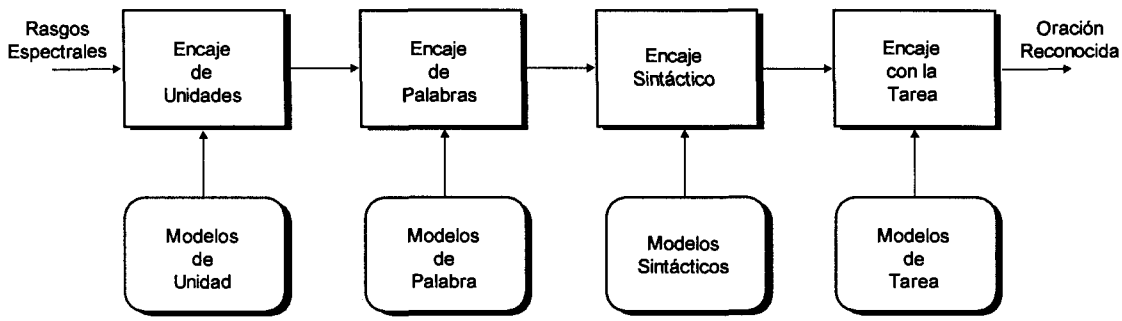

Figura 3.15. Diagrama de bloques de un reconocedor de habla continua modular.

En los últimos años se ha realizado un progreso significativo en el desarrollo de los algoritmos de búsqueda [JEL94] [GUP95]. Los ejemplos más destacados son:

1. Algoritmo de búsqueda en haz de una sola pasada (*One-Pass Beam Search*). Para la parte del conocimiento que puede integrarse dentro de la red de estados finitos, el problema de búsqueda se suele resolver encontrando el mejor camino posible a través de la red. Un algoritmo de búsqueda exhaustiva por paso, como el algoritmo de Viterbi, resulta muy costoso en términos de tiempo de procesamiento y capacidad de almacenaje. En este algoritmo solamente un pequeño conjunto de todas las hipótesis parciales plausibles (palabras) que caigan dentro del haz es evaluado. Diversas técnicas

como son el uso de árboles léxicos y la búsqueda anticipada de fonemas se han implementado con objeto de reducir el número de hipótesis durante la búsqueda. La anchura del haz, que determina los costes computacionales y el posible número de errores producto de la poda en el árbol de búsqueda, se establece de manera empírica, siendo dependiente de la tarea y del modelado empleado.

Hay que hacer notar que la solución proporcionada por este método de búsqueda no es la óptima, puesto que proporciona la probabilidad de una secuencia de estados individuales dentro del modelo compuesto y no la probabilidad de emisión total de la secuencia de modelos que constituye la mejor palabra. En la práctica, sin embargo, se observa que la probabilidad calculada con este algoritmo presenta la propiedad de dominio, que consiste en que el camino así hallado concentra la mayor parte de la probabilidad total [MER91].

2. Algoritmos de búsqueda heurística basados en el A* [NIL80] y decodificación con pila (*stack decoding*) [PAU91], [PAU92], [GOP95a], [GOP95b]. Teniendo en cuenta que la señal de voz porta información lingüística de alguna manera localizada, no todos los eventos lingüísticos están activos y necesitan ser evaluados en todo momento. Con el fin de asimilar esta propiedad se puede usar una estrategia de búsqueda del tipo el mejor primero (*best first*). Esta búsqueda se suele implementar mediante el uso de una pila que mantiene para cada instante de tiempo una lista ordenada con las hipótesis parciales. Entonces, la mejor hipótesis de la pila se intenta extender a una pequeña lista de palabras, elegida de acuerdo a la bondad de los encajes en los niveles acústico y gramatical. Una de las ventajas de este método es que los modelos de lengua a largo plazo pueden integrarse de forma natural en la búsqueda. Para controlar el crecimiento exponencial de la pila se suelen emplear estrategias de búsqueda basadas en el algoritmo A*. De esta manera se emplea, no sólo la estimación del coste de un camino hasta el punto alcanzado, sino también una evaluación del camino que falta por recorrer.

Otro ejemplo de algoritmo basado en el A*, pero aplicado a la búsqueda en un grafo constituido por elementos silábicos lo encontramos en [GUP88]. La gran ventaja de este método es el menor tamaño de la red como consecuencia del menor número de elementos básicos (sílabas en este caso).

El algoritmo del enrejado en árbol (*tree-trellis*) [SOO91] es una forma eficaz de controlar el tamaño de la pila al mantener todas las posibles opciones de continuación en un árbol y recombinarlas con las hipótesis parciales hacia atrás, que presentes en la pila, se han ido clasificado en un segundo árbol de búsqueda.

3. Estrategias de decisión con múltiples pasadas. En oposición a los algoritmos tradicionales de búsqueda de izquierda a derecha (una sola pasada), este tipo de algoritmos realiza una primera pasada con el fin de establecer las hipótesis iniciales y en vueltas posteriores se completan de forma progresiva [MUR93], [ALL93] [AUB94]. Los métodos muti-paso, como el mencionado anteriormente, se suelen diseñar con el fin de proporcionar las mejores N hipótesis (oraciones) [SCH90], [SCH92]. Este paradigma de búsqueda resulta ideal cuando se desea integrar múltiples fuentes de conocimiento a diferentes niveles de abstracción y que no son fácilmente integrables.

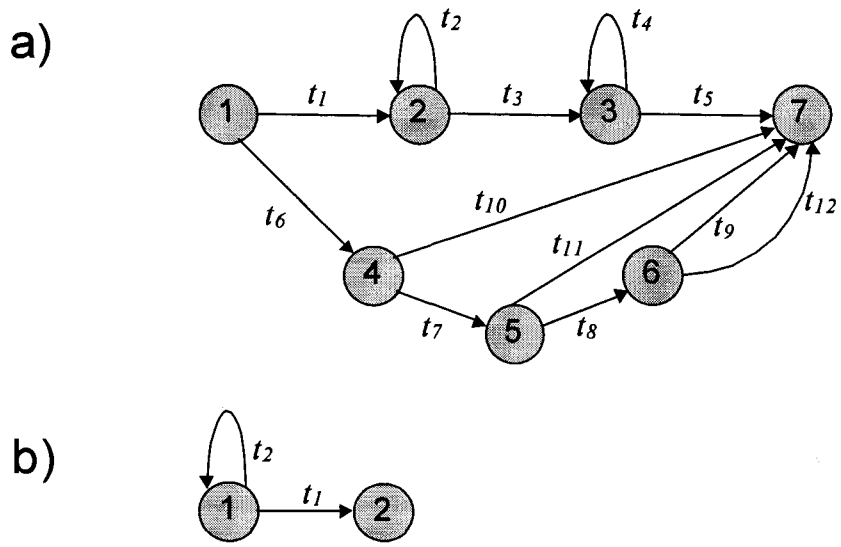

Figura 3.16. Estructura de un modelo oculto de Markov correspondiente a un fonema. a).- Estructura estándar. b).- Estructura reducida.

Con objeto de mejorar la flexibilidad, se pueden emplear modelos acústicos más sencillos para producir un retículo con segmentos (*segment lattice*) o con fonemas (*phone lattice*), en la primera pasada de la búsqueda. Los modelos léxicos y de lengua pueden además incorporarse para generar un retículo de palabras (*word lattice*). La

familia de métodos que se conoce con el nombre de búsqueda progresiva (*progressive search*), es una forma muy interesante de aunar diferentes niveles de conocimiento de una manera sistemática con el fin de mejorar la calidad del reconocimiento.

Para reducir el número de hipótesis posibles durante el proceso de búsqueda se puede usar un procedimiento de encaje rápido conocido como *fast match* (Figura 3.16) sobre una red reducida con modelos acústicos simplificados.

Algunas modificaciones a los algoritmos de búsqueda anteriores con objeto de mejorar las prestaciones de los respectivos sistemas pasan por añadir restricciones de tipo temporal a los diferentes segmentos de la red de búsqueda [BUS87], [SOO89].

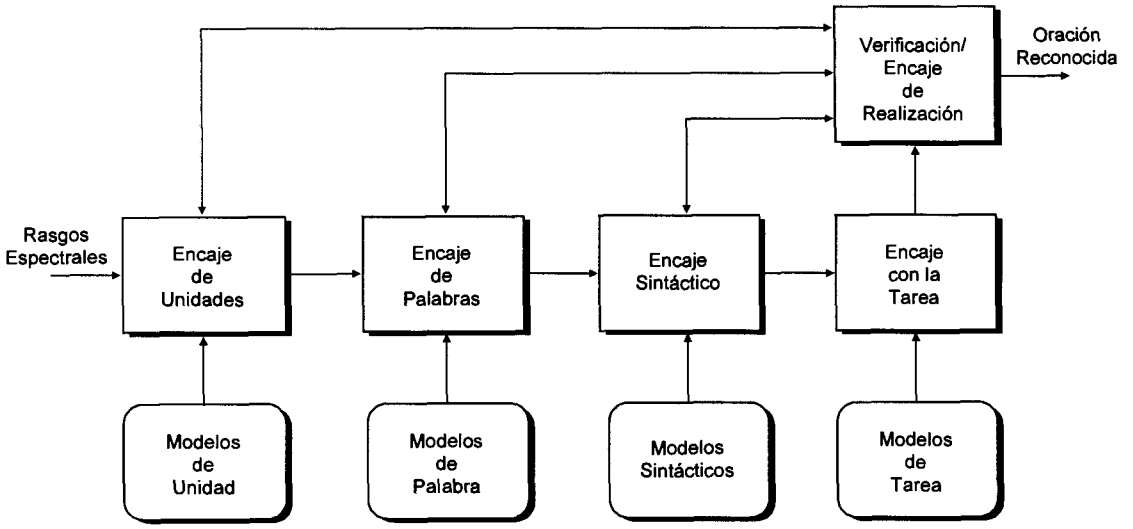

Figura 3.17. Diagrama de bloques de un reconocedor de habla continua modular con integración de conocimiento.

La incorporación de diferentes fuentes de información (a veces incompatibles entre sí) implica también la necesidad de establecer un modelo que permita manejar diferentes hipótesis para cada una de ellas. Una posibilidad es la que aparece en la (Figura 3.17), en la cual las soluciones parciales de los cuatro módulos se integran con el fin de determinar la solución global.

La combinación de este tipo de aproximación con estrategias de verificación de realizaciones resulta ser una manera bastante flexible de diseñar sistemas apropiados para tratar con grandes vocabularios [RAB96].

4. ANÁLISIS COMPARATIVO DE LOS MÉTODOS DE EXTRACCIÓN DE PARÁMETROS

4.1. INTRODUCCIÓN

El propósito del presente capítulo es analizar el comportamiento de los diferentes tipos de extractores de rasgos, que están basados en las técnicas presentadas en el capítulo 3. Para ello lo primero de todo es definir el marco de trabajo, con objeto de poder comprender la validez de los comentarios presentes en los siguientes apartados.

El sistema de reconocimiento de voz empleado en todas las pruebas es un reconocedor de palabras aisladas con las siguientes características:

a).- Extractor de parámetros variable para cada experimento. En concreto se probaron 35 diferentes modelos de extracción de rasgos basados en LPC (Tabla 4.1), FFT (Tabla 4.2) y PLP (Tabla 4.3).

b).- Cuantificador vectorial de 256 índices.

c).- Modelos ocultos de Markov discretos. Se utilizó un modelo por cada palabra. Todos los modelos son de 6 estados con transiciones al mismo estado y al siguiente, únicamente.

d).- La detección de principio y fin de palabra se realiza a través de estimaciones de energía derivadas del proceso de extracción paramétrica.

e).- Sin modelo léxico ni gramatical.

Para la batería de ensayos se utilizaron dos bases de datos de palabras:

1. **Base de datos de mandatos para un teléfono extendido**. Presenta las órdenes necesarias para controlar por voz las funciones de un teléfono de servicios extendidos. Permite estudiar la variabilidad introducida por la diversidad de hablantes, así como, por los diferentes tipos de entonación empleados por un mismo locutor.

2. **Base de datos del proyecto IVORY** [IVO.html]. Contiene las órdenes más habituales para un mando de control de dirección así como para marcado de dígitos. Permite analizar los efectos del ruido en los sistemas de reconocimiento, así como, la variabilidad introducida por la multiplicidad de

hablantes y por el dialectalismo propio de hablantes de una segunda lengua.

La base de datos de mandatos para un teléfono extendido (Tabla 4.4) consta de 27 palabras del castellano y se registraron en condiciones silenciosas. El número de hablantes es de 26 y la distribución entre voces masculinas y femeninas es equitativa.

Cada locutor realizó una serie completa de toda la lista de palabras empleando un modo de entonación característico (enunciativo, interrogativo o exclamativo). En total cada hablante realizó un total de 6 series con los 3 tipos de entonación mencionados, es decir, 2 series por cada modo de entonación.

Código Parametrización	Descripción	Número y Tipo de Parámetros
LPC-1	LPC-Coeficientes *a*	16 coeficientes
LPC-2		16 coeficientes + 16 coeficientes
LPC-3	LPC-Cepstrum	10 coeficientes
LPC-4		10 coeficientes + 10 coeficientes delta (5 tramas)
LPC-5	LPC-Bandas Energía	20 bandas
LPC-6		20 bandas + 20 bandas delta (5 tramas)
LPC-7		20 bandas + 20 bandas diferencia (2 tramas)
LPC-8	LPC-Bandas Energía Normalizadas	20 bandas
LPC-9		20 bandas + 20 bandas delta (5 tramas)
LPC-10		20 bandas + 20 bandas diferencia (2 tramas)
LPC-11	LPC-Bandas Energía Cepstrum	10 coeficientes
LPC-12		10 coeficientes + 10 coeficientes delta (5 tramas)

Tabla 4.1. Conjunto de experimentos basados en parametrizadores LPC.

Código Parametrización	Descripción	Número y Tipo de Parámetros
FFT-1	FFT-Bandas Energía	20 bandas
FFT-2		20 bandas + 20 bandas delta (5 tramas)
FFT-3		20 bandas + 20 bandas diferencia (2 tramas)
FFT-4	FFT-Bandas Energía Normalizadas	20 bandas
FFT-5		20 bandas + 20 bandas delta (5 tramas)
FFT-6		20 bandas + 20 bandas diferencia (2 tramas)
FFT-7	FFT-Cepstrum	10 coeficientes
FFT-8		10 coeficientes + 10 coeficientes delta (5 tramas)
FFT-9		10 coeficientes + 10 coeficientes delta + 10 coeficientes delta-delta (5 tramas)
FFT-10		10 coeficientes + 10 coeficientes delta + 10 coeficientes delta-delta (5 tramas) + energía trama actual + energía trama diferencia (2 tramas)

Tabla 4.2. Conjunto de experimentos basados en parametrizadores FFT.

Código Parametrización	Descripción	Número y Tipo de Parámetros
PLP-1	PLP	10 coeficientes
PLP-2		10 coeficientes + 10 coeficientes delta (5 tramas)
PLP-3		10 coeficientes + 10 coeficientes delta (5 tramas) + 10 coeficientes delta-delta
PLP-4		10 coeficientes + 10 coeficientes delta (5 tramas) + 10 coeficientes delta-delta + energía trama actual + energía trama diferencia (2 tramas)
PLP-5		10 coeficientes + 10 coeficientes delta (5 tramas). Sin normalización de energía
PLP-6		10 coeficientes + 10 coeficientes delta (5 tramas). Sin cepstrum
PLP-7		10 coeficientes + 10 coeficientes delta (5 tramas). Sin parte perceptual
PLP-8	P-Cepstrum	10 coeficientes
PLP-9		10 coeficientes + 10 coeficientes delta (5 tramas)
PLP-10	RASTA-PLP	10 coeficientes
PLP-11		10 coeficientes + 10 coeficientes delta (5 tramas)
PLP-12	RASTA-P-Cepstrum	10 coeficientes
PLP-13		10 coeficientes + 10 coeficientes delta (5 tramas)

Tabla 4.3. Conjunto de experimentos basados en parametrizadores PLP.

Código	Palabra	Código	Palabra
(1)	Aceptar	(15)	Mensaje
(2)	Adelante	(16)	Menú
(3)	Atrás	(17)	Nueve
(4)	Cancelar	(18)	Ocho
(5)	Cero	(19)	Recibir
(6)	Cinco	(20)	Repetir
(7)	Cuatro	(21)	Seis
(8)	Dos	(22)	Servicio
(9)	Enviar	(23)	Siete
(10)	Establecer	(24)	Teléfono
(11)	Fax	(25)	Texto
(12)	Información	(26)	Tres
(13)	Internet	(27)	Uno
(14)	Marcar		

Tabla 4.4. Conjunto de palabras de la base de datos de mandatos para un teléfono extendido.

Por su parte, la base de datos del proyecto IVORY consta de 30 palabras del inglés (Tabla 4.5) realizadas en 3 condiciones ambientales diferentes:

1. Ambiente silencioso.

2. Ambiente ruidoso. El ruido se introdujo en el ambiente durante las grabaciones

usando una fuente que simulaba una sala de videojuegos, siendo el origen principal, un juego de *rallies*. El micrófono estaba situado a 1,5 m de la fuente de ruido y los niveles de ruido medidos con un sonómetro estándar se mantuvieron alrededor de los 90-95 dB.

3. Ambiente silencioso pero con ruido inducido en el hablante, es decir, de habla con *stress*. La generación se realizó por medio de la introducción del ruido del punto anterior a través de un sistema de auriculares de tipo abierto.

El número de hablantes es de 40 y la distribución entre voces masculinas y femeninas es equitativa. Cada locutor realizó una serie completa de toda la lista de palabras para cada uno de los ambientes descritos.

Código	Palabra	Código	Palabra
(1)	Double	(16)	One
(2)	Down	(17)	On
(3)	Eight	(18)	Right
(4)	End	(19)	Seven
(5)	Five	(20)	Six
(6)	Four	(21)	Split
(7)	Go	(22)	Stand
(8)	Hit	(23)	Stop
(9)	Jump	(24)	Ten
(10)	Last	(25)	Three
(11)	Left	(26)	Turn
(12)	Next	(27)	Two
(13)	Nine	(28)	Up
(14)	No	(29)	Yes
(15)	Off	(30)	Zero

Tabla 4.5. Conjunto de palabras de la base de datos del proyecto IVORY.

En ambos casos los locutores tenían al castellano como lengua materna y sus edades estaban comprendidas entre los 18 y los 50 años, siendo la edad media de 22,5.

La instrumentación empleada para el registro fue un micrófono de calibración con lóbulo hipercardioide y una tarjeta de sonido Fiji Turtle Beach con CODEC de 18 bits. Las grabaciones se realizaron a 22.050 Hz y posteriormente se remuestraron vía *software* a 11025 Hz. Posteriormente se procedió al cortado y etiquetado de los registros

en ficheros de palabra individuales. El procesamiento se realizó por lotes siguiendo diversos esquemas de entrenamiento.

Para la base de datos con órdenes para un teléfono extendido, se establecieron un total de 10 experimentos por cada uno de los 35 extractores paramétricos (Tabla 4.6). En 6 de estos esquemas [*(a)- (e)*], se emplean realizaciones con entonación enunciativa para entrenar el sistema. Los otros 6, además, se entrenan con realizaciones de modos interrogativo y exclamativo.

Código	Número Locutores		Entonación	
	Hombres	Mujeres	Enunciativa	Interrogativa y Exclamativa
(a)	1	1	✓	×
(b)	3	3	✓	×
(c)	6	0	✓	×
(d)	0	6	✓	×
(e)	6	6	✓	×
(f)	1	1	✓	✓
(g)	3	3	✓	✓
(h)	6	0	✓	✓
(i)	0	6	✓	✓
(j)	6	6	✓	✓

Tabla 4.6. Esquema de los 10 entrenamientos [*(a)-(j)*] para la base de datos de mandatos para un teléfono extendido.

Dentro de cada conjunto de 6 pruebas, se realiza la inclusión progresiva de más hablantes durante el entrenamiento. Inicialmente se prueba con un solo locutor de cada género para pasar a (3+3) y por último a 6 hablantes masculinos y 6 femeninos. Por último, se incluyen 2 ensayos en los que emplean únicamente hablantes de un mismo género.

Las pruebas que se realizan sobre el conjunto de 10 entrenamientos son de dos tipos (Tabla 4.7). En las primeras, columna etiquetada como *(1)*, se chequean solamente las realizaciones producidas con modo enunciativo, mientras que en las segundas se prueba con los otros tipos de entonación. La lista de locutores de estas pruebas incluye 7 hablantes masculinos y 7 femeninos, todos diferentes a los que se han empleado durante la construcción del sistema.

Código	Número Locutores		Entonación	
	Hombres	Mujeres	Enunciativa *(1)*	Interrogativa y Exclamativa *(2)*
(a)	7	7	✓	✓
(b)	7	7	✓	✓
(c)	7	7	✓	✓
(d)	7	7	✓	✓
(e)	7	7	✓	✓
(f)	7	7	✓	✓
(g)	7	7	✓	✓
(h)	7	7	✓	✓
(i)	7	7	✓	✓
(j)	7	7	✓	✓

Tabla 4.7. Esquema de los 2 tipos de test [*(1), (2)*] que se aplican sobre los 10 modelos de entrenamiento [*(a)-(j)*] empleados con la base de datos de mandatos para un teléfono extendido. Cada uno de estas configuraciones se aplica a los distintos esquemas de extracción de parámetros.

Por su parte, para la base de datos del proyecto IVORY, se establecieron un total de 6 experimentos por parametrizador (Tabla 4.8). En 3 de estos esquemas [*(a)- (c)*], se usan para construir el sistema, realizaciones registradas en condiciones silenciosas. En el resto además se emplean las realizaciones de habla con *stress*.

Dentro de cada conjunto de 3 pruebas, se realiza también como en el caso anterior, la inclusión progresiva de más hablantes durante el entrenamiento. Inicialmente se prueba con un solo locutor masculino y femenino para llegar posteriormente, a 8 hablantes masculinos y 8 femeninos.

Las pruebas que se realizan sobre el conjunto de 6 entrenamientos son de tres tipos (Tabla 4.9):

1. Realizaciones silenciosas.
2. Realizaciones de habla con *stress*.
3. Realizaciones ruidosas.

La lista de locutores de estas pruebas la forman 12 hablantes masculinos y 12

femeninos, todos diferentes a los que se han empleado durante la construcción del sistema asociado.

Código	Número Locutores		Entorno de Grabación		
	Hombres	Mujeres	Silencio	Ruido Inducido	Ruido
(a)	1	1	✓	×	×
(b)	4	4	✓	×	×
(c)	8	8	✓	×	×
(d)	1	1	✓	✓	×
(e)	4	4	✓	✓	×
(f)	8	8	✓	✓	×

Tabla 4.8. Cuadro de los 6 entrenamientos [*(a)- (f)*] para la base de datos del proyecto IVORY. Cada uno de estas configuraciones se aplica a los distintos esquemas de extracción de parámetros.

Código	Número Locutores		Entorno de Grabación		
	Hombres	Mujeres	Silencio *(1)*	Ruido Inducido *(2)*	Ruido *(3)*
(a)	12	12	✓	✓	✓
(b)	12	12	✓	✓	✓
(c)	12	12	✓	✓	✓
(d)	12	12	✓	✓	✓
(e)	12	12	✓	✓	✓
(f)	12	12	✓	✓	✓

Tabla 4.9. Cuadro de los 3 tipos de test [*(1)- (3)*] que se aplican sobre los 6 modelos de entrenamiento [*(a)- (f)*] empleados con la base de datos del proyecto IVORY.

En ambos casos, también se obtienen los resultados de la evaluación de cada uno de los sistemas anteriormente citados con los datos empleados en su construcción.

En cada experimento, el tipo de resultados que se van a contrastar es:

a).- Tasa de reconocimiento global para un conjunto de palabras. Indica el porcentaje de aciertos durante el reconocimiento respecto al número de palabras totales testadas. Esta medida se conoce también como tasa de error de palabras o WER (*Word Error Rate*) y se define [COL97]:

$$WER = 100 \cdot \frac{N_{inserciones} + N_{borrados} + N_{sustituciones}}{N} \tag{4.1}$$

Como puede comprobarse el origen de los errores lo encontramos en tres causas:

1. Errores de inserción. Se producen cuando el sistema detecta palabras no existentes.
2. Errores de borrado. Cuando el reconocedor no detecta palabras presentes en el discurso.
3. Errores de sustitución. La palabra detectada, aun cuando existe, no es la correcta.

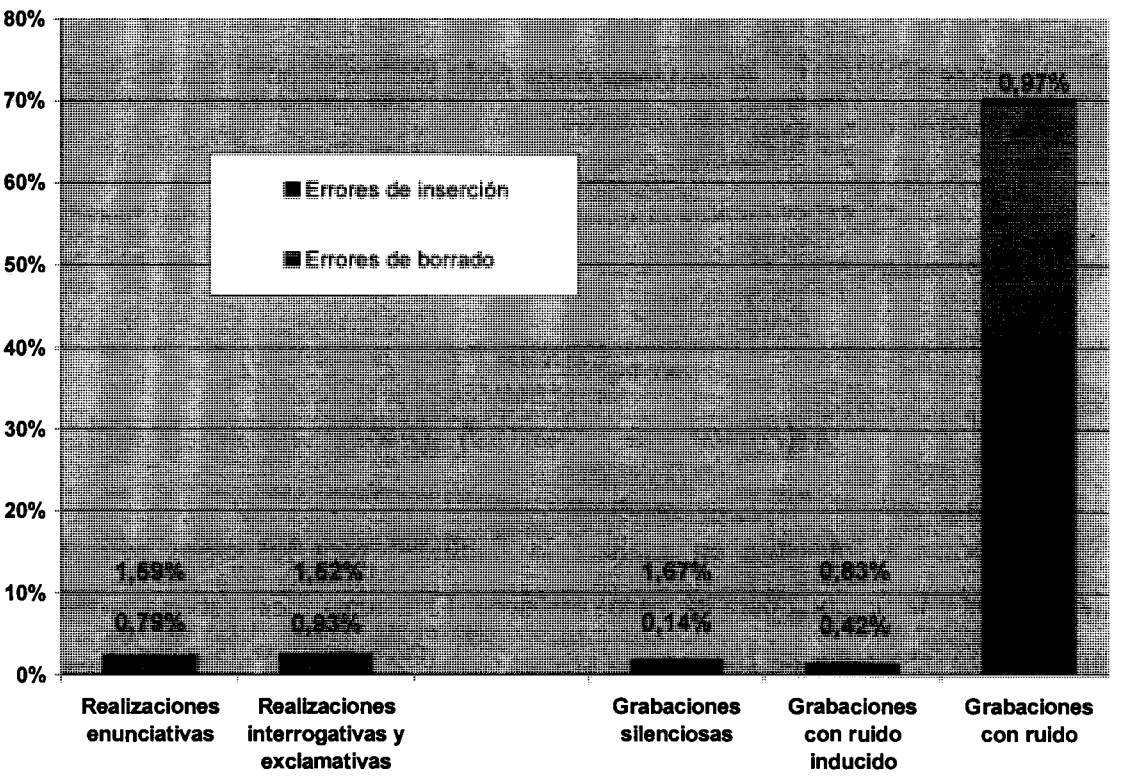

Figura 4.1. Tasa de error para los errores de inserción y borrado presentes en los 5 grupos de test como consecuencia de fallos en la detección de principio y fin de palabra.

En nuestro caso, los errores de inserción y borrado son debidos exclusivamente al procedimiento de detección de principio y fin y no los vamos a tener en cuenta en los siguientes análisis puesto que su participación en el error es la misma para los diferentes tipos de parametrización.

Sin embargo, esto no quiere decir que su contribución al error total sea despreciable (Figura 4.1). Los errores de inserción, para el caso de reconocimiento de palabras aisladas, se producen por la presencia de pequeños ruidos en las grabaciones como por ejemplo soplidos en el micrófono. En experimentos de laboratorio pueden ser fácilmente eliminados pero aportan una mayor naturalidad y realismo a éstos. Los errores de borrado se producen como consecuencia de la falta de contraste en los niveles de energía durante los periodos de voz y silencio, siendo este fenómeno extremadamente relevante en condiciones de ruido.

b).- Índice de fiabilidad global. Proporciona una medida muy útil para determinar la calidad del proceso de reconocimiento [BOU96]. Un sistema puede tener una tasa de reconocimiento mayor que otro, pero si éste último puede determinar cuáles son los casos erróneos o al menos los dudosos, existe la posibilidad de eliminar estos errores utilizando otras estrategias tales como pedir confirmación al usuario o aplicar un segundo análisis de reserva.

Para nuestros experimentos emplearemos dos medidas de fiabilidad basadas en la siguiente fórmula:

$$Fiabilidad = \frac{100 \cdot \log\left(\frac{P(HMM\ 2^a\ mejor\ probabilidad)}{P(HMM\ mejor\ probabilidad)}\right)}{\log\left(P(HMM\ mejor\ probabilidad) \cdot P(HMM\ 2^a\ mejor\ probabilidad)\right)} \qquad (4.2)$$

La primera consistirá en analizar la fiabilidad para los casos correctamente clasificados y la segunda para los casos erróneos. El sistema ideal debería proporcionar un valor alto (alrededor de 100) para el primer caso y bajo para el segundo (alrededor de 0). Hay que tener en cuenta que los valores de probabilidad para un modelo serán números comprendidos en el intervalo *(0.0, 1.0)* y que el logaritmo proporcionará un resultado negativo siempre.

c).- Tasa de reconocimiento individual de las palabras de un conjunto. Es la medida correspondiente al apartado a) pero aplicada a cada palabra de forma individual. Resulta interesante porque permite comprobar cuales son los casos que más afectan al mal funcionamiento del sistema.

d).- Índices de fiabilidad individuales para cada palabra. Es la medida correspondiente al apartado b) pero aplicada a cada palabra de forma individual.

e).- Tasa de confusión entre las palabras de un mismo grupo. Permite conocer aquellas palabras que reunidas en el mismo grupo tienden a producir más errores. Con esta información se puede refinar la construcción de los conjuntos de palabras empleando sinónimos.

4.2. ANÁLISIS DE EXPERIMENTOS BASADOS EN LPC

Presentamos en esta sección los resultados para el conjunto de 12 parametrizadores LPC-1 al LPC-12, incluyendo los casos de rasgos estáticos, es decir, aquellos calculados únicamente sobre la trama actual y dinámicos (calculados en parte empleando las informaciones de tramas consecutivas).

Dividimos los métodos de este grupo en cinco apartados:

1. LPC-Coeficientes *a*.
2. LPC-Cepstrum.
3. LPC-Bandas Energía.
4. LPC-Bandas Energía Normalizadas.
5. LPC-Bandas Energía Cepstrum.

4.2.1. LPC-Coeficientes a

4.2.1.1. Parámetros estáticos

4.2.1.1.1. Experimento LPC-1

a).- Procedimiento de cálculo:

1. Algoritmo celosía adaptativa de gradiente con 16 etapas.
2. Rutina de Levinson-Durbin para obtener coeficientes del filtro *a*, partiendo de los coeficientes de reflexión *h(i)*.

b).- Resultados de reconocimiento:

- Presenta las peores tasas de reconocimiento de todo el conjunto de parametrizadores analizado en el presente estudio para las dos bases de datos.
- Los resultados son peores para el caso de la base de datos del proyecto IVORY (8,03%, 8,93%, 4,40%) que para los datos del teléfono extendido (37,52%, 31,63%).
- Dentro de cada grupo de palabras, los resultados son muy parecidos con independencia del tipo de tramas empleadas en las fases de entrenamiento y test.
- Cuando los conjuntos de entrenamiento y test coinciden, las tasas son extremadamente pobres (71,44%, 46,27%). También resulta notorio en este apartado la caída de los porcentajes que se produce por la inclusión de nuevos hablantes. Este mismo fenómeno se produce cuando se entrena con modos interrogativo y exclamativo, así como con ambiente de ruido inducido.

c).- Índices de fiabilidad:

- Presenta los peores índices de todo el conjunto de parametrizadores analizado.
- El valor de los índices de fiabilidad para los aciertos está siempre por debajo de 6,0.

- Incluso en las grabaciones con ambiente ruidoso, se presenta la aparente contradicción de que el índice para los aciertos es menor que para las realizaciones erróneamente clasificadas.
- La separación entre los índices de acierto y fallo es prácticamente nula, siendo muy parecidos en todos los casos.
- Solamente para los casos de prueba con los datos del entrenamiento, con un número mínimo de hablantes, modos de entonación y ambientes, los resultados son buenos <0,0- 30,40>, <0,0- 28,59>.

4.2.1.2. Parámetros dinámicos

4.2.1.2.1. Experimento LPC-2

a).- Procedimiento de cálculo:

1. Algoritmo celosía adaptativa de gradiente con 16 etapas. Se obtienen 16 coeficientes PARCOR *h(i)*.
2. Rutina de Levinson-Durbin para obtener los coeficientes del filtro *a(i)*, partiendo de los coeficientes de reflexión *h(i)*.
3. Concatenación del vector de coeficientes de la trama actual *t* y el correspondiente al periodo anterior *t-1*.

b).- Resultados de reconocimiento:

- Después del anterior, proporciona las peores tasas de reconocimiento de todo el conjunto de parametrizadores analizado. El único caso en el que otro parametrizador resulta ser peor es en el test con grabaciones ruidosas, aunque por muy poco.
- Aún así, los resultados son mucho mejores que para el caso anterior. En el caso de la base de datos del teléfono extendido (51,67%, 50,22%) y para los datos del proyecto IVORY se obtiene: (21,03%, 18,53%, 9,00%).
- La tendencia es por lo demás muy similar a la del parametrizador estático en cuanto a su comportamiento.

c).- Índices de fiabilidad:

- Presenta junto con el LPC-1 los peores índices de todo el conjunto de parametrizadores analizado.
- Al igual que en ese caso, los valores para los casos de acierto y fallos son prácticamente idénticos en todos los casos; siendo la separación casi nula.

4.2.2. LPC-Cepstrum

4.2.2.1. Parámetros estáticos

4.2.2.1.1. Experimento LPC-3

a).- Procedimiento de cálculo:

1. Algoritmo celosía adaptativa de gradiente con 16 etapas.
2. Rutina de Levinson-Durbin para obtener coeficientes del filtro *a*, partiendo de los coeficientes de reflexión *h(i)*.
3. Cálculo de 10 coeficientes cepstrales derivados de los coeficientes *a(i)*.

b).- Resultados de reconocimiento:

- Los resultados de reconocimiento para este caso caen dentro de la zona templada, siendo mejores que para el caso dinámico anterior (LPC-2). Para la base de datos del teléfono extendido tenemos como mejores resultados: (70,00%, 69,85%) y para los datos del proyecto IVORY: (39,43%, 34,00%, 8,07%).
- La única excepción la constituye el test con grabaciones ruidosas, donde este método ocupa la segunda plaza después del LPC-1.
- El comportamiento de los sistemas es similar con independencia de que se incluya mayor o menor variedad de realizaciones durante la construcción del sistema (habla con modo exclamativo/interrogativo y habla con *stress*).
- La capacidad del reconocedor para absorber mayor cantidad de hablantes es ahora mucho mayor que en los casos anteriores, aunque siempre existe una ligera pérdida según se emplea más material. Esto aparece reflejado claramente en el test con los datos del entrenamiento. Este mismo efecto se observa para todo el resto de modelos de extracción de características, en mayor o menor medida.

c).- Índices de fiabilidad:

- Mejoran respecto a los dos casos anteriores (LPC-1 y LPC-2) aunque mantienen las mismas tendencias.
- Aunque los valores para los casos acertados son algo mayores, por lo general los índices de fiabilidad para los errores también aumentan.

4.2.2.2. Parámetros dinámicos

4.2.2.2.1. Experimento LPC-4

a).- Procedimiento de cálculo:

1. Algoritmo celosía adaptativa de gradiente con 16 etapas.
2. Rutina de Levinson-Durbin para obtener coeficientes del filtro *a*, partiendo de los coeficientes de reflexión *h(i)*.
3. Cálculo de 10 coeficientes cepstrales derivados de los coeficientes *a(i)*.
4. Cálculo de 10 coeficientes Δ-cepstrum a partir de 5 tramas (*t-4, t-3, t-2, t-1 y t*).

b).- Resultados de reconocimiento:

- Respecto al parametrizador estático (LPC-3), las tasas en este caso son algo mejores para la base de datos del teléfono extendido (73,74%, 75,52%) y peores para el otro caso (39,03%, 33,17%, 6,47%), especialmente en ambiente ruidoso.

c).- Índices de fiabilidad:

- Se mantienen en los términos del apartado anterior.
- Solamente con la base de datos del proyecto IVORY, se percibe una pérdida en la discriminación.

4.2.3. LPC-Bandas Energía

4.2.3.1. Parámetros estáticos

4.2.3.1.1. Experimento LPC-5

a).- Procedimiento de cálculo:

1. Algoritmo celosía adaptativa de gradiente con 16 etapas.
2. Rutina de Levinson-Durbin para obtener coeficientes del filtro *a*, partiendo de los coeficientes de reflexión *h(i)*.
3. Cálculo de la función de transferencia del filtro.
4. Cálculo de 20 bandas en frecuencia siguiendo escalas de Mel.

b).- Resultados de reconocimiento:

- Aporta los mejores resultados en todas las pruebas del conjunto de test de entre los parametrizadores estáticos basados en técnicas LPC: (85,89%, 85,83%) y (50,43%, 49,50%, 17,23%).

c).- Índices de fiabilidad:

- Se advierte una clara mejoría respecto a los parametrizadores anteriores.
- Con los datos de la base del teléfono extendido presentan los mejores resultados de entre los parametrizadores estáticos LPC: (<2,21- 10,50>, <2,90-10,45>). Incluso con el test de realizaciones con modo enunciativo presentan los segundos mejores resultados.
- Con la base de datos del proyecto IVORY, los resultados son más discretos en comparación a otros métodos que veremos a continuación.

4.2.3.2. Parámetros dinámicos

4.2.3.2.1. Experimento LPC-6

a).- Procedimiento de cálculo:

1. Algoritmo celosía adaptativa de gradiente con 16 etapas.
2. Rutina de Levinson-Durbin para obtener coeficientes del filtro *a*, partiendo de los coeficientes de reflexión *h(i)*.
3. Cálculo de la función de transferencia del filtro.
4. Cálculo de 20 bandas en frecuencia siguiendo escalas de Mel.
5. Cálculo de 20 bandas en frecuencia Δ a partir de 5 tramas (*t-4, t-3, t-2, t-1 y t*).

b).- Resultados de reconocimiento:

- Presenta los mejores resultados de reconocimiento para todos los casos del grupo de test para las dos bases de datos (excepto con ambiente ruidoso cuando se entrena conjuntamente con realizaciones limpias y con habla con *stress*).
- Al igual que la mayoría de los parametrizadores, el entrenar con mayor variedad de realizaciones proporciona mejores resultados. La única excepción en este caso la proporciona también el caso ruidoso.
- Destaca su comportamiento cuando se entrena utilizando tan solo dos hablantes (uno masculino y otro femenino). En esta situación obtenemos las siguientes tasas de reconocimiento: (72,25%, 73,67%), (82,22%, 82,07%) para la base de datos del teléfono extendido y (34,07%, 27, 63%, 15,83%), (35,87%, 34,30%, 9,93%) para los datos de proyecto IVORY.
- Cuando el entrenamiento se realiza tomando el máximo de hablantes posible sus resultados son: (88,11, 89,48%), (93,07, 94,30%) y (60,63%, 48,57%, 17,30%), (62,47%, 61,57%, 15,90%).

c).- Índices de fiabilidad:

- Su tendencia sigue el comportamiento de las tasas de reconocimiento.
- Presenta los mejores índices en todos los casos del grupo de test con excepción de las grabaciones ruidosas.
- La diferencia entre índices de aciertos y fallos es mejor para la base de datos

del teléfono extendido (<1,82- 14,01>, <1,96- 13,93>) que para el otro grupo (<4,42- 9,49>, <3,65- 9,37>, <3,30, 3,19>).

4.2.3.2.2. Experimento LPC-7

a).- Procedimiento de cálculo:

1. Algoritmo celosía adaptativa de gradiente con 16 etapas.
2. Rutina de Levinson-Durbin para obtener coeficientes del filtro *a*, partiendo de los coeficientes de reflexión *h(i)*.
3. Cálculo de la función de transferencia del filtro.
4. Cálculo de 20 bandas en frecuencia siguiendo escalas de Mel.
5. Concatenación del vector de coeficientes de la trama actual *t* y el correspondiente al periodo anterior *t-1*.

b).- Resultados de reconocimiento:

- Sus porcentajes de reconocimiento se sitúan por encima de los del experimento LPC-5 (caso estático asociado) pero por debajo del parametrizador anterior.
- Con la base de datos del proyecto IVORY proporciona los segundos mejores resultados para los casos de test con realizaciones silenciosas y con habla con *stress* (57,57%, 57,17%) respectivamente.

c).- Índices de fiabilidad:

- Los índices asociados a este experimento siguen los patrones de casos anteriores sin destacar en ninguna de las categorías.

4.2.4. LPC-Bandas Energía Normalizadas

4.2.4.1. Parámetros estáticos

4.2.4.1.1. Experimento LPC-8

a).- Procedimiento de cálculo:

1. Algoritmo celosía adaptativa de gradiente con 16 etapas.
2. Rutina de Levinson-Durbin para obtener coeficientes del filtro *a*, partiendo de los coeficientes de reflexión *h(i)*.
3. Cálculo de la función de transferencia del filtro.
4. Cálculo de 20 bandas en frecuencia normalizadas en el rango *[0.0, 1.0]* siguiendo escalas de Mel.

b).- Resultados de reconocimiento:

- En comparación al experimento LPC-5 (sin normalización de parámetros), los resultados se sitúan ligeramente por debajo en todos los casos.

c).- Índices de fiabilidad:

- Los índices de fiabilidad asociados son también peores que los del parametrizador más parecido a éste (LPC-5).

4.2.4.2. Parámetros dinámicos

4.2.4.2.1. Experimento LPC-9

a).- Procedimiento de cálculo:

1. Algoritmo celosía adaptativa de gradiente con 16 etapas.
2. Rutina de Levinson-Durbin para obtener coeficientes del filtro *a*, partiendo de los coeficientes de reflexión *h(i)*.
3. Cálculo de la función de transferencia del filtro.

4. Cálculo de 20 bandas en frecuencia normalizadas en el rango *[0.0, 1.0]* siguiendo escalas de Mel.

5. Cálculo de 20 bandas en frecuencia Δ a partir de 5 tramas (*t-4, t-3, t-2, t-1 y t*).

b).- Resultados de reconocimiento:

- Los resultados de este modelo son mejores que los de su versión estática LPC-8 y que los del experimento LPC-5 (estático sin normalización). Sin embargo son peores que para el caso LPC-6 (dinámico sin normalización).
- Se observa además cómo este parametrizador amplifica la tendencia de pérdida en las tasas de reconocimiento, que se produce en el cambio estático a dinámico para los casos LPC-5 al LPC-6.
- La excepción la constituyen las grabaciones con ambiente ruidoso cuando se entrena con realizaciones silenciosas y de habla con *stress*. Con este marco de trabajo, los resultados que se alcanzan son los segundos mejores del grupo (21,80%).

c).- Índices de fiabilidad:

- En general se mantiene el patrón seguido por las tasas de reconocimiento, aunque en alguna situación los índices sean mejores que los de otros parametrizadores.
- Lo más destacable son sus índices para el caso ruidoso <2,70- 3,49>. A pesar de su pobreza discriminatoria constituyen el mejor caso.

4.2.4.2.2. Experimento LPC-10

a).- Procedimiento de cálculo:

1. Algoritmo celosía adaptativa de gradiente con 16 etapas.
2. Rutina de Levinson-Durbin para obtener coeficientes del filtro *a*, partiendo de los coeficientes de reflexión *h(i)*.
3. Cálculo de la función de transferencia del filtro.

4. Cálculo de 20 bandas en frecuencia normalizadas en el rango *[0.0, 1.0]* siguiendo escalas de Mel.

5. Concatenación del vector de coeficientes de la trama actual *t* y el correspondiente al periodo anterior *t-1*.

b).- Resultados de reconocimiento:

- Sus resultados con la excepción del caso del test con modo enunciativo (88,89%) son peores que los resultados asociados con los parametrizadores de tipo dinámico de este grupo.

- Para la base de datos del proyecto IVORY las tasas se sitúan en valores muy similares a las de su parametrizador estático LPC-8, aunque ligeramente superiores.

c).- Índices de fiabilidad:

- Presenta índices similares a los últimos experimentos sin destacar en ninguna categoría.

4.2.5. LPC-Bandas Energía Cepstrum

4.2.5.1. Parámetros estáticos

4.2.5.1.1. Experimento LPC-11

a).- Procedimiento de cálculo:

1. Algoritmo celosía adaptativa de gradiente con 16 etapas.
2. Rutina de Levinson-Durbin para obtener coeficientes del filtro *a*, partiendo de los coeficientes de reflexión *h(i)*.
3. Cálculo de la función de transferencia del filtro.
4. Cálculo de 20 bandas en frecuencia siguiendo escalas de Mel.
5. Cálculo de 10 coeficientes siguiendo la expresión del cálculo de coeficientes MFCC.

b).- Resultados de reconocimiento:

- Las tasas de reconocimiento para los datos del teléfono extendido son claramente los peores de entre los parametrizadores basados en medidas de bandas de energía (74,22%, 74,85%).
- Para la base de datos del proyecto IVORY son sin embargo muy similares a los experimentos LPC-5 y LPC-8 (estáticos). Incluso para el caso silencioso son mejores que éstos (50,73%, 48,13%, 14,77%).

c).- Índices de fiabilidad:

- Para la aplicación del teléfono extendido los índices de fiabilidad son peores que para el experimento LPC-5 pero mejores que los del supuesto LPC-8.
- Con la otra base de datos, los índices son especialmente buenos. De hecho, con la excepción del caso ruidoso <4,04- 3,52>, constituyen los mejores (<5,54-11,32>, <5,50-10,81>).

4.2.5.2. Parámetros dinámicos

4.2.5.2.1. Experimento LPC-12

a).- Procedimiento de cálculo:

1. Algoritmo celosía adaptativa de gradiente con 16 etapas.
2. Rutina de Levinson-Durbin para obtener coeficientes del filtro *a*, partiendo de los coeficientes de reflexión *h(i)*.
3. Cálculo de la función de transferencia del filtro.
4. Cálculo de 20 bandas en frecuencia siguiendo escalas de Mel.
5. Cálculo de 10 coeficientes siguiendo la expresión del cálculo de coeficientes MFCC.
6. Cálculo de 10 coeficientes Δ-cepstrum a partir de 5 tramas (*t-4, t-3, t-2, t-1 y t*).

b).- Resultados de reconocimiento:

- La mejora experimentada en las tasas de reconocimiento para los datos del teléfono extendido es notoria. Aunque son ligeramente inferiores a los del supuesto LPC-6, están por encima de los de experimento LPC-9. Se convierte por tanto en los mejores segundos resultados para los experimentos basados en LPC (89,78%, 91,15%).
- Para la base de datos del proyecto IVORY sucede algo similar. Quedan encuadrados entre los experimentos LPC-6 y LPC-8.
- La mejor parte de este modelo lo constituye los resultados del caso ruidoso. Presenta con 23,23% la mejor tasa de todo este grupo de experimentos.

c).- Índices de fiabilidad:

- Los índices de fiabilidad, en comparación con el experimento LPC-11 (estático), no siguen las mismas tendencias.
- Para la base de datos del teléfono, los índices de este parametrizador están en línea con los del LPC-11.

- Con los datos del proyecto IVORY, los índices son bastante peores si exceptuamos el caso ruidoso, que ahora está ente los mejores <2,95- 3,57>.

4.3. ANÁLISIS DE EXPERIMENTOS BASADOS EN FFT

Presentamos en esta sección los resultados para el conjunto de 10 parametrizadores FFT-1 al FFT-10, incluyendo algunos de ellos rasgos dinámicos.

Dividimos los métodos de este grupo en tres apartados:

1. FFT-Bandas Energía.
2. FFT-Bandas Energía Normalizadas.
3. FFT-Cepstrum.

4.3.1. FFT-Bandas Energía

4.3.1.1. Parámetros estáticos

4.3.1.1.1. Experimento FFT-1

a).- Procedimiento de cálculo:

1. Algoritmo transformada discreta de Fourier (radix 2 diezmado en el tiempo) con 256 muestras y desplazando 128.
2. Cálculo de 20 bandas en frecuencia siguiendo escalas de Mel.

b).- Resultados de reconocimiento:

- Las tasas de reconocimiento para este experimento están entre las peores para buen número de situaciones, en concreto para las pruebas con modo interrogativo/exclamativo (84,52%) y con ambientes silencioso y ruidoso (48,17%, 5,63%).
- En el resto de situaciones los resultados están lejos de ser los mejores.

c).- Índices de fiabilidad:

- Los índices de fiabilidad para las dos bases de datos están más cerca de los peores resultados que de los mejores. Aun así, no presenta valores especialmente destacables.

4.3.1.2. Parámetros dinámicos

4.3.1.2.1. Experimento FFT-2

a).- Procedimiento de cálculo:

1. Algoritmo transformada discreta de Fourier (radix 2 diezmado en el tiempo) con 256 muestras y desplazando 128.
2. Cálculo de 20 bandas en frecuencia siguiendo escalas de Mel.
3. Cálculo de 20 bandas en frecuencia Δ a partir de 5 tramas (*t-4, t-3, t-2, t-1 y t*).

b).- Resultados de reconocimiento:

- Destaca en este experimento la gran mejora experimentada en casi todos los apartados respecto al caso anterior FFT-1.
- Para la base de datos del teléfono extendido, los resultados con el test de realizaciones con modo enunciativo son los mejores de este grupo (92,78%). Para el otro supuesto la mejora es considerable pero los resultados lo sitúan en la zona media (89,11%).
- Para la base de datos del proyecto IVORY presenta los mejores resultados de todos los experimentos en este grupo, con la única excepción de las grabaciones ruidosas, donde aparecen unos resultados bastante malos (64,90%, 69,33%, 5,83%).

c).- Índices de fiabilidad:

- Los índices de fiabilidad de este parametrizador se encuentran entre los mejores para todos los casos, si exceptuamos las realizaciones ruidosas. En las pruebas con la base de datos del proyecto IVORY sus resultados son: (<3,80- 7,96>, <3,71- 8,21>, <1,28- 0,90>).
- Merece una mención especial los índices de las pruebas con la base de datos de teléfono extendido (<1,50- 10,92>, <1,93- 10,28>), donde sus resultados son los mejores de entre todos los experimentos basados en FFT.

4.3.1.2.2. Experimento FFT-3

a).- Procedimiento de cálculo:

1. Algoritmo transformada discreta de Fourier (radix 2 diezmado en el tiempo) con 256 muestras y desplazando 128.
2. Cálculo de 20 bandas en frecuencia siguiendo escalas de Mel.
3. Concatenación del vector de coeficientes de la trama actual *t* y el correspondiente al periodo anterior *t-1*.

b).- Resultados de reconocimiento:

- Sus resultados de reconocimiento son peores que para el experimento anterior.
- Incluso para las pruebas con modo exclamativo/interrogativo y ambiente ruidoso las tasas son peores que para el parametrizador estático asociado (FFT-1).

c).- Índices de fiabilidad:

- Los índices de fiabilidad son también malos en todos los casos.
- Excepto para las grabaciones silenciosas, los valores están entre los peores de todo este grupo de parametrizadores.

4.3.2. FFT-Bandas Energía Normalizadas

4.3.2.1. Parámetros estáticos

4.3.2.1.1. Experimento FFT-4

a).- Procedimiento de cálculo:

1. Algoritmo transformada discreta de Fourier (radix 2 diezmado en el tiempo) con 256 muestras y desplazando 128.
2. Cálculo de 20 bandas en frecuencia normalizadas en el rango *[0.0, 1.0]* y siguiendo escalas de Mel.

b).- Resultados de reconocimiento:

- Junto al experimento FFT-1 comparte los peores resultado de este grupo de experimentos. En concreto las tasas para los ensayos con modo enunciativo y habla con *stress* presentan los peores valores (84,89%, 45,13%)
- Solamente el caso ruidoso está por encima de la media (17,17%).

c).- Índices de fiabilidad:

- Los índices de fiabilidad para este parametrizador son en general malos. En especial destaca su pobre comportamiento con la base de datos del teléfono extendido (<1,42- 7,10>, <2,10- 7,32>).
- Para las pruebas con ambiente ruidoso proporciona los segundos mejores índices (<2,09- 2,73>).

4.3.2.2. Parámetros dinámicos

4.3.2.2.1. Experimento FFT-5

a).- Procedimiento de cálculo:

1. Algoritmo transformada discreta de Fourier (radix 2 diezmado en el tiempo) con 256 muestras y desplazando 128.

2. Cálculo de 20 bandas en frecuencia normalizadas en el rango *[0.0, 1.0]* y siguiendo escalas de Mel.

3. Cálculo de 20 bandas en frecuencia Δ a partir de 5 tramas (*t-4, t-3, t-2, t-1 y t*).

b).- Resultados de reconocimiento:

- Supone una mejora sustancial respecto al experimento anterior especialmente para los casos relacionados con la base de datos del teléfono extendido. Sus resultados con estas grabaciones están entre los dos mejores (92,00%, 93,93%).
- Con la otra base de datos la mejora es más discreta (54,67%, 53,83%, 17,90%), si exceptuamos el caso de habla con *stress*. A pesar de que las tasas para este último punto no son especialmente buenas, la progresión es significativa.

c).- Índices de fiabilidad:

- Con la base de datos del teléfono extendido los índices de fiabilidad son los segundos mejores (<1,40- 9,87>, <1,90- 10,16>).
- Con los otros datos, sus índices no destacan en ningún capítulo y son muy parecidos a los del experimento FFT-4.

4.3.2.2.2. Experimento FFT-6

a).- Procedimiento de cálculo:

1. Algoritmo transformada discreta de Fourier (radix 2 diezmado en el tiempo) con 256 muestras y desplazando 128.

2. Cálculo de 20 bandas en frecuencia normalizadas en el rango *[0.0, 1.0]* y siguiendo escalas de Mel.

3. Concatenación del vector de coeficientes de la trama actual *t* y el correspondiente al periodo anterior *t-1*.

b).- Resultados de reconocimiento:

- En comparación con los parametrizadores más parecidos (FFT-4 y FFT5), no

aporta nada especialmente relevante en las pruebas con la base de datos telefónica. Sus resultados son mejores que para el supuesto estático y peores que para el dinámico.

- Pésimos resultados en los casos de habla limpia y con *stress* (49,00%, 49,33%). También con grabaciones ruidosas su comportamiento es peor que el de los ensayos FFT-4 y FFT-5 (13,19%).

c).- Índices de fiabilidad:

- Al igual que sucedía con las tasas de reconocimiento, los índices de fiabilidad para las pruebas con habla limpia y con stress son de los peores que encontramos en este punto (<2,51- 4,92>, <2,52- 4,90>).

4.3.3. FFT-Cepstrum

4.3.3.1. Parámetros estáticos

4.3.3.1.1. Experimento FFT-7

a).- Procedimiento de cálculo:

1. Algoritmo transformada discreta de Fourier (radix 2 diezmado en el tiempo) con 256 muestras y desplazando 128.
2. Cálculo de 20 bandas en frecuencia siguiendo escalas de Mel.
3. Cálculo de 10 coeficientes siguiendo la expresión del cálculo de coeficientes MFCC.

b).- Resultados de reconocimiento:

- Destaca su buen comportamiento con la base de datos del proyecto IVORY, donde sus resultados (56,73%, 54,83%, 16,70%), se encuentran por encima de los parametrizadores estáticos FFT-1 y FFT-4 (salvo en el caso ruidoso). Aunque globalmente no son los mejores, superan a los experimentos dinámicos FFT-5 y FFT6 en situaciones de habla limpia y con *stress*.
- Para la otra base de datos, las tasas no destacan en absoluto aunque tampoco desmerecen.

c).- Índices de fiabilidad:

- Las conclusiones son análogas al apartado de reconocimiento.

4.3.3.2. Parámetros dinámicos

4.3.3.2.1. Experimento FFT-8

a).- Procedimiento de cálculo:

1. Algoritmo transformada discreta de Fourier (radix 2 diezmado en el tiempo) con 256 muestras y desplazando 128.

2. Cálculo de 20 bandas en frecuencia siguiendo escalas de Mel.
3. Cálculo de 10 coeficientes siguiendo la expresión del cálculo de coeficientes MFCC.
4. Cálculo de 10 coeficientes Δ-cepstrum a partir de 5 tramas (*t-4, t-3, t-2, t-1 y t*).

b).- Resultados de reconocimiento:

- Para la base de datos del proyecto IVORY, este esquema de parametrización constituye el mejor modelo en promedio (62,87%, 61,97%, 26,40%). Aunque sus resultados son algo peores que para el FFT-2, en las pruebas con realizaciones silenciosas y con *stress*, su resultado para el ensayo con ambiente ruidoso resulta ser el mejor.
- Con los otros datos, las tasas de reconocimiento se encuentran en la zona media respecto a los otros parametrizadores dinámicos del grupo.

c).- Índices de fiabilidad:

- Reproducen el mismo patrón de comportamiento que los porcentajes de reconocimiento.
- Los valores obtenidos con la base de datos del proyecto IVORY son los mejores en promedio (<2,91- 748>, <3,16- 7,57>, <2,64- 3,68>). Solamente para la situación de habla con stress, los índices son algo peores que para el experimento FFT-2.

4.3.3.2.2. Experimento FFT-9

a).- Procedimiento de cálculo:

1. Algoritmo transformada discreta de Fourier (radix 2 diezmado en el tiempo) con 256 muestras y desplazando 128.
2. Cálculo de 20 bandas en frecuencia siguiendo escalas de Mel.
3. Cálculo de 10 coeficientes siguiendo la expresión del cálculo de coeficientes MFCC.

4. Cálculo de 10 coeficientes Δ-cepstrum a partir de 5 tramas (*t-4, t-3, t-2, t-1 y t*).

5. Cálculo de 10 coeficientes Δ-(Δ-cepstrum) a partir de 5 tramas (*t-4, t-3, t-2, t-1 y t*).

b).- Resultados de reconocimiento:

- Los resultados de este experimento son muy similares a los del anterior FFT-8.
- Con la excepción del test con las grabaciones de modos exclamativo e interrogativo, donde se aprecia una ligera mejoría, el resto de casos muestra una sutil pérdida como consecuencia de añadir al vector de parámetros las características delta-delta.

c).- Índices de fiabilidad:

- Son muy similares al caso FFT-8, pero en ningún caso mejores (salvo para la prueba con modo enunciativo).

4.3.3.2.3. Experimento FFT-10

a).- Procedimiento de cálculo:

1. Algoritmo transformada discreta de Fourier (radix 2 diezmado en el tiempo) con 256 muestras y desplazando 128.
2. Cálculo de 20 bandas en frecuencia siguiendo escalas de Mel.
3. Cálculo de 10 coeficientes siguiendo la expresión del cálculo de coeficientes MFCC.
4. Cálculo de 10 coeficientes Δ-cepstrum a partir de 5 tramas (*t-4, t-3, t-2, t-1 y t*).
5. Cálculo de 10 coeficientes Δ-(Δ-cepstrum) a partir de 5 tramas (*t-4, t-3, t-2, t-1 y t*).
6. Concatenación con energía de la trama y energía diferencia calculada a partir de 2 tramas (*t-1* y *t*).

b).- Resultados de reconocimiento:

- Representa una continuación de los resultados del caso anterior.
- Sólo con las pruebas con modo enunciativo presenta una ligera mejoría (89,52%).
- Dentro de la base de datos del proyecto IVORY, sus tasas caen espectacularmente, siendo el caso ruidoso especialmente malo (59,57%, 52,00%, 9,50%).

c).- Índices de fiabilidad:

- Este parametrizador no aporta ninguna mejora en este capítulo frente al modelo anterior.

4.4. ANÁLISIS DE EXPERIMENTOS BASADOS EN PLP

Por último, presentamos en esta sección los resultados para el conjunto de 13 parametrizadores basados en PLP: PLP-1 al PLP-13.

Dividiremos los métodos de este grupo en cuatro apartados:

1. PLP.
2. P-Cepstrum.
3. RASTA-PLP.
4. RASTA-P-Cepstrum.

4.4.1. PLP

4.4.1.1. Parámetros estáticos

4.4.1.1.1. Experimento PLP-1

a).- Procedimiento de cálculo:

1. Algoritmo de predicción lineal perceptual con ventana de 10 ms. Se obtienen 10 coeficientes LPC-cepstrum.

b).- Resultados de reconocimiento:

- Comparado con los otros parametrizadores estáticos de este grupo presenta un comportamiento bueno con las dos bases de datos (83,22%, 87,04%), (51,27%, 45,60%, 12,20%).
- Mejora frente a éstos (PLP-8, PLP-10, PLP-12) en las pruebas con habla silenciosa y con *stress*, así como cuando el ensayo se realiza con grabaciones con modo enunciativo.

c).- Índices de fiabilidad:

- Siguen un razonamiento similar que para las tasas de reconocimiento.
- Destacan los índices para las realizaciones silenciosas (<2,54- 5,58>).

4.4.1.2. Parámetros dinámicos

4.4.1.2.1. Experimento PLP-2

a).- Procedimiento de cálculo:

1. Algoritmo de predicción lineal perceptual con ventana de 10 ms. Se obtienen 10 coeficientes LPC-cepstrum.
2. Cálculo de 10 coeficientes Δ a partir de 5 tramas (*t-4, t-3, t-2, t-1 y t*).

b).- Resultados de reconocimiento:

- En comparación al caso estático (PLP-1) presenta una clara mejoría con todos los conjuntos de prueba (89,85%, 93,56%), (58,37%, 53,83%, 18,00%).

- En comparación con otros parametrizadores dinámicos, destaca su comportamiento en el test con realizaciones silenciosas, donde es el mejor y con realizaciones ruidosas.

c).- Índices de fiabilidad:

- Con la base de datos del teléfono extendido presenta los segundos mejores índices (<1,44- 10,38>, <1,55- 9,93>).

- Con el otro conjunto de grabaciones proporciona también buenos índices (<2,75- 6.03>, <2,32- 5,98>, <2,47- 3,25>), sobre todo con grabaciones silenciosas y ruidosas.

4.4.1.2.2. Experimento PLP-3

a).- Procedimiento de cálculo:

1. Algoritmo de predicción lineal perceptual con ventana de 10 ms. Se obtienen 10 coeficientes LPC-cepstrum.
2. Cálculo de 10 coeficientes Δ a partir de 5 tramas (*t-4, t-3, t-2, t-1 y t*).
3. Cálculo de 10 coeficientes Δ^2 a partir de 5 tramas (*t-4, t-3, t-2, t-1 y t*).

b).- Resultados de reconocimiento:

- Proporciona las mejores tasas de reconocimiento dentro de este apartado, tanto en los casos de realizaciones ruidosas (57,00%) como de habla con *stress* (57,00%).

- Las tasas del resto de casos se encuentran en límites similares a los que podemos encontrar para el modelo PLP-2.

c).- Índices de fiabilidad:

- Son peores que para el experimento anterior en todos las situaciones, situándose en una zona media dentro de los experimentos basados en PLP.

4.4.1.2.3. Experimento PLP-4

a).- Procedimiento de cálculo:

1. Algoritmo de predicción lineal perceptual con ventana de 10 ms. Se obtienen 10 coeficientes LPC-cepstrum.
2. Cálculo de 10 coeficientes Δ a partir de 5 tramas (*t-4, t-3, t-2, t-1 y t*).
3. Cálculo de 10 coeficientes Δ^2 a partir de 5 tramas (*t-4, t-3, t-2, t-1 y t*).
4. Concatenación con energía de la trama y energía diferencia calculada a partir de 2 tramas (*t-1* y *t*).

b).- Resultados de reconocimiento:

- Para la base de datos telefónica, este modelo proporciona mejores resultados que los aportados por los esquemas PLP-2 y PLP-3 (91,19%, 93,96%). El resultado para realizaciones con entonación exclamativa e interrogativa es el segundo mejor.
- Con la otra base de datos, los resultados de este parametrizador empeoran si los comparamos con los modelos anteriores.

c).- Índices de fiabilidad:

- No aporta ninguna mejora respecto a PLP-2 y PLP-3.

4.4.1.2.4. Experimento PLP-5

a).- Procedimiento de cálculo:

1. Algoritmo de predicción lineal perceptual con ventana de 10 ms. Se obtienen 10 coeficientes LPC-cepstrum sin normalización de energía.
2. Cálculo de 10 coeficientes Δ a partir de 5 tramas (*t-4, t-3, t-2, t-1 y t*).

b).- Resultados de reconocimiento:

- Comparado con el parametrizador más parecido (PLP-2) el test con grabaciones con modo interrogativo/exclamativo arroja una ligera pérdida (92,26%), pero por el contrario, en el caso con modo enunciativo (92,26%) es el mejor de entre todos los parametrizadores.
- En cuanto a la otra base de datos (proyecto IVORY), los resultados de habla con *stress* son algo mejores (54,27%), pero los de habla normal peores (52,70%). Los resultados de las grabaciones ruidosas son francamente mucho peores (7,70%).

c).- Índices de fiabilidad:

- Los índices de fiabilidad son peores para este modelo que para el PLP-2 en todas las situaciones.
- Resulta especialmente importante la pérdida de las capacidades discriminativas en los ensayos con la base de datos del proyecto IVORY (<0,85- 1,95>, <0,91- 2,29 >, <0.95- 1-23>).

4.4.1.2.5. Experimento PLP-6

a).- Procedimiento de cálculo:

1. Algoritmo de predicción lineal perceptual con ventana de 10 ms. Se obtienen 10 coeficientes LPC.
2. Cálculo de 10 coeficientes Δ a partir de 5 tramas (*t-4, t-3, t-2, t-1 y t*).

b).- Resultados de reconocimiento:

- Se experimenta una pérdida significativa en todos los resultados de reconocimiento respecto al caso PLP-2.
- Dentro de la base de datos telefónica la pérdida es apreciable (81,22%, 80,22%, 92,52%), situándose por debajo incluso del caso estático más similar (PLP-1).
- Para la otra base de datos, los resultados son los peores dentro de este conjunto

de experimentos (25,77%, 24,97%, 3,33%).

c).- Índices de fiabilidad:

- Claramente empeoran con respecto a los ensayos PLP-2.
- Sus índices de fiabilidad se encuadran dentro de la línea general trazada por el experimento anterior (PLP-5).

4.4.1.2.6. Experimento PLP-7

a).- Procedimiento de cálculo:

1. Algoritmo de predicción lineal perceptual sin parte perceptual y con ventana de 10 ms. Se obtienen 10 coeficientes LPC-cepstrum.
2. Cálculo de 10 coeficientes Δ a partir de 5 tramas (*t-4, t-3, t-2, t-1 y t*).

b).- Resultados de reconocimiento:

- Presenta una doble faceta dependiendo de la base de datos considerada.
- Con la base de datos del teléfono extendido los resultados están entre los mejores de este grupo (91,56%, 94,00%).
- Sin embargo, las tasas de reconocimiento son mucho peores con la base de datos del proyecto IVORY (47,47%, 45,03%, 7,33%). De hecho estos resultados son peores que los de su parametrizador estático asociado (PLP-1)

c).- Índices de fiabilidad:

- Son relativamente buenos para la base de datos telefónica, aunque por debajo de los mejores resultados vistos hasta hora (PLP-2).
- Por el contrario, los índices de fiabilidad para el otro grupo de grabaciones están entre los peores (<0,94- 1,90>, <1,08- 2,18>, <0,77- 0,92>).

4.4.2. P-Cepstrum

4.4.2.1. Parámetros estáticos

4.4.2.1.1. Experimento PLP-8

a).- Procedimiento de cálculo:

1. Algoritmo de predicción lineal perceptual con ventana de 10 ms. Se obtienen 10 coeficientes LPC-cepstrum. No incluye la parte de predicción lineal del algoritmo.

b).- Resultados de reconocimiento:

- Presenta los peores resultados de reconocimiento de todos los parametrizadores basados en PLP para la base de datos telefónica (62,52%, 54,85%) y el segundo peor para la otra (27,02%, 27,83%, 7,93%).

c).- Índices de fiabilidad:

- Presenta los peores índices de fiabilidad de todos los parametrizadores basados en PLP para la base de datos telefónica (<2,38- 5,33>, <2,08- 4,81>).
- Para los datos del proyecto IVORY los índices de fiabilidad son mucho peores que para el experimento PLP-1.

4.4.2.2. Parámetros dinámicos

4.4.2.2.1. Experimento PLP-9

a).- Procedimiento de cálculo:

1. Algoritmo de predicción lineal perceptual con ventana de 10 ms. Se obtienen 10 coeficientes LPC-cepstrum. No incluye la parte de predicción lineal del algoritmo.
2. Cálculo de 10 coeficientes Δ a partir de 5 tramas (*t-4, t-3, t-2, t-1 y t*).

b).- Resultados de reconocimiento:

- Las tasas de reconocimiento mejoran bastante con relación al experimento anterior PLP-8 (caso estático asociado).
- No obstante, los resultados están siempre por debajo del experimento PLP-2 y en algunas ocasiones por debajo de la versión estática de éste (PLP-1).

c).- Índices de fiabilidad:

- Los índices de fiabilidad se sitúan más cerca del experimento PLP-1 que de su variante dinámica PLP-2.

4.4.3. RASTA-PLP

4.4.3.1. Parámetros estáticos

4.4.3.1.1. Experimento PLP-10

a).- Procedimiento de cálculo:

1. Algoritmo de predicción lineal perceptual y filtrado RASTA con ventana de 10 ms. Se obtienen 10 coeficientes LPC-cepstrum.

b).- Resultados de reconocimiento:

- Presenta unos resultados muy similares al experimento PLP-1 aunque generalmente algo por debajo.
- Solamente proporciona un mejor resultado, aunque muy pequeño, en las pruebas realizadas con grabaciones ruidosas (12,47%) y con modo enunciativo (84,26%).

c).- Índices de fiabilidad:

- Presenta resultados muy similares al modelo PLP-1.
- Las únicas mejoras, casi imperceptibles, las encontramos en los ensayos de habla con *stress*, ruido y modo interrogativo/exclamativo.

4.4.3.2. Parámetros dinámicos

4.4.3.2.1. Experimento PLP-11

a).- Procedimiento de cálculo:

1. Algoritmo de predicción lineal perceptual y filtrado RASTA con ventana de 10 ms. Se obtienen 10 coeficientes LPC-cepstrum.
2. Cálculo de 10 coeficientes Δ a partir de 5 tramas (*t-4, t-3, t-2, t-1 y t*).

b).- Resultados de reconocimiento:

- Experimenta una clara mejoría respecto a su parametrizador estático (PLP-10).
- En comparación al modelo PLP-2, es mejor en los ensayos de habla con *stress* (55,40%) y con modo interrogativo/exclamativo (93,67%).
- Sin embargo es peor para las situaciones con ruido (16,73%).

c).- Índices de fiabilidad:

- Los índices de fiabilidad son los mejores de entre todos los parametrizadores de este grupo para la base de datos telefónica (<1,52- 10,59>, <1,73- 10,16>). No obstante no se separan mucho de los obtenidos para el experimento PLP-2
- Para la otra base de datos, los índices son algo inferiores a los del caso PLP-2.

4.4.4. RASTA-P-Cepstrum

4.4.4.1. Parámetros estáticos

4.4.4.1.1. Experimento PLP-12

a).- Procedimiento de cálculo:

1. Algoritmo de predicción lineal perceptual y filtrado RASTA con ventana de 10 ms. Se obtienen 10 coeficientes LPC-cepstrum. No incluye la parte de predicción lineal del algoritmo.

b).- Resultados de reconocimiento:

- Presenta unos resultados parecidos a los aportados por el experimento PLP-8 y siempre bastante por debajo de los que presenta el modelo PLP-10, es decir, la eliminación de la parte de predicción lineal resulta clave.

c).- Índices de fiabilidad:

- Los índices de fiabilidad reflejan el mismo comportamiento visto para el reconocimiento.
- En todos los casos los índices son ligeramente mejores que para el experimento PLP-10.

4.4.4.2. Parámetros dinámicos

4.4.4.2.1. Experimento PLP-13

a).- Procedimiento de cálculo:

1. Algoritmo de predicción lineal y filtrado RASTA perceptual con ventana de 10 ms. Se obtienen 10 coeficientes LPC-cepstrum. No incluye la parte de predicción lineal del algoritmo.
2. Cálculo de 10 coeficientes Δ a partir de 5 tramas (*t-4, t-3, t-2, t-1 y t*).

b).- Resultados de reconocimiento:

- Presenta un paralelismo bastante acusado con relación al comportamiento del caso estático anterior.
- Los resultados para este caso, aunque son superiores a la versión estática del parametrizador, se encuentran más cerca del experimento PLP-9 que del PLP-11.

c).- Índices de fiabilidad:

- Los índices de fiabilidad siguen el mismo comportamiento para la base de datos del proyecto IVORY.
- En cambio, para los otros datos, los índices son los mejores que encontramos dentro de los esquemas de parametrización basados en PLP (<3,14- 6,74>, <3,64- 7,23>).

4.5. ANÁLISIS DE TASAS DE ERROR Y FIABILIDAD POR PALABRA

4.5.1. Estudio de las tasas de error por palabra

En este apartado vamos a revisar de manera breve los resultados de reconocimiento de forma individualizada, es decir, desde el punto de vista de los elementos básicos o palabras. No analizaremos todos y cada uno de los modelos de parametrización vistos en el apartado anterior, sino que, nos referiremos únicamente a los casos con mejores tasas de reconocimiento para cada uno de los cinco grandes grupos de test:

1. Test de realizaciones con modo enunciativo (Base de datos del teléfono extendido).
2. Test de realizaciones con modo exclamativo/interrogativo (Base de datos del teléfono extendido).
3. Test de realizaciones silenciosas (Base de datos de proyecto IVORY.).
4. Test de realizaciones de habla con *stress* (Base de datos de proyecto IVORY.).
5. Test de realizaciones ruidosas (Base de datos de proyecto IVORY.).

En la Tabla 4.10, Tabla 4.11 y Tabla 4.12 aparecen recogidas las tasas de error para las palabras con mayor número de fallos. Cada tabla incluye los resultados de un grupo de técnicas de extracción de parámetros estudiados en los apartados anteriores: LPC, FFT y PLP. Los resultados completos de estos experimentos, incluyendo las tablas de confusión entre palabras, pueden verse en el Apéndice B.

La primera impresión que sugiere la inspección de las tablas es que las palabras que concentran la mayor parte de los errores son un grupo reducido. Esto además se verifica para todos los casos de test, sea cual sea el experimento.

Especialmente significativo es el caso de los ensayos con realizaciones ruidosas,

puesto que aquí encontramos series de vocablos que sistemáticamente son confundidos y contribuyen muy negativamente a los resultados finales.

Base de datos del teléfono extendido				Base de datos del proyecto IVORY					
Modo Enunciativo		***Modo Interrogativo y Exclamativo***		***Grabaciones Silenciosas***		***Grabaciones con Ruido Inducido***		***Grabaciones Ruidosas***	
LPC-6		LPC-6		LPC-6		LPC-6		LPC-12	
Palabra	**Tasa de Error**	**Palabra**	**Tasa de Error**	**Palabra**	**Tasa de Error**	**Palabra**	**Tasa de Error**	**Palabra**	**Tasa de Error**
Nueve	25%	Aceptar	24%	Ten	80%	Ten	67%	Double	100%
Cero	18%	Cero	15%	Seven	59%	Double	63%	Hit	100%
Aceptar	15%	Fax	13%	Turn	59%	Seven	63%	Jump	100%
Recibir	15%	Recibir	13%	Double	55%	Turn	63%	Seven	100%
Tres	15%			End	55%	Two	63%	Six	100%
								Split	100%
								Stand	100%
								Stop	100%
								Ten	100%
								Turn	100%

Tabla 4.10. Tasas de error para las palabras de diversos experimentos basados en LPC.

Base de datos del teléfono extendido				Base de datos del proyecto IVORY					
Modo Enunciativo		***Modo Interrogativo y Exclamativo***		***Grabaciones Silenciosas***		***Grabaciones con Ruido Inducido***		***Grabaciones Ruidosas***	
FFT-2		FFT-5		FFT-2		FFT-2		FFT-8	
Palabra	**Tasa de Error**	**Palabra**	**Tasa de Error**	**Palabra**	**Tasa de Error**	**Palabra**	**Tasa de Error**	**Palabra**	**Tasa de Error**
Uno	22%	Cero	24%	Ten	71%	End	67%	Double	100%
Cancelar	18%	Tres	15%	Down	63%	Last	67%	Down	100%
Cero	18%	Texto	13%	Left	55%	Go	59%	Hit	100%
Fax	18%	Aceptar	11%	Last	50%	Two	59%	Seven	100%
		Fax	11%	Eight	46%	One	55%	Six	100%
		Nueve	11%	Go	46%	Turn	55%	Split	100%
		Seis	11%					Stop	100%
								Three	100%

Tabla 4.11. Tasas de error para las palabras de diversos experimentos basados en FFT.

Base de datos del teléfono extendido				Base de datos del proyecto IVORY					
Modo Enunciativo		*Modo Interrogativo y Exclamativo*		*Grabaciones Silenciosas*		*Grabaciones con Ruido Inducido*		*Grabaciones Ruidosas*	
PLP-5 (351)		PLP-7 (371)		PLP-2 (321)		PLP-3 (331)		PLP-3 (331)	
Palabra	**Tasa de Error**	**Palabra**	**Tasa de Error**	**Palabra**	**Tasa de Error**	**Palabra**	**Tasa de Error**	**Palabra**	**Tasa de Error**
Aceptar	25%	Cero	29%	Ten	71%	End	75%	Hit	100%
Cero	22%	Aceptar	18%	Down	63%	Turn	75%	On	100%
Seis	22%	Mensaje	17%	Last	63%	Last	71%	Standt	100%
Uno	22%	Tres	8%	Turn	63%	Stand	59%	Stop	100%
Mensaje	18%			Hit	59%	Yes	59%	Three	100%
Dos	15%			Eight	55%				
				Go	55%				
				On	55%				

Tabla 4.12. Tasas de error para las palabras de diversos experimentos basados en PLP.

Respecto a la confusión entre palabras destacaremos varios aspectos:

a).- Se puede observar como muchos de los casos referidos en el apartado a) se confunden en gran número con otra determinada palabra. Algunos ejemplos pueden ser ***aceptar-cancelar***, ***double-down*** y ***ten-end***. Resulta curioso, sin embargo, comprobar como la dirección del error no es siempre la misma.

b).- Las otras palabras tienden a repartir más el error entre el resto, aunque esta tendencia no sigue un criterio aleatorio. Esto quiere decir que cuando un término se confunde con más de uno, tiende a repetirse en los demás experimentos al menos una parte de ellos. Un ejemplo puede ser ***cero*** y ***servicio*** o ***stop*** y ***up***.

4.5.2. Estudio de los índices de fiabilidad por palabra

Al igual que hacíamos antes, en las tablas Tabla 4.13, Tabla 4.14 y Tabla 4.15 recogemos los índices de fiabilidad de las mejores y peores palabras, agrupadas por tipo de experimento.

Los resultados muestran una clasificación que varía de ++ (mejor índice conjunto) a -- (peor índice). Este marcador conjunto se calcula a partir de los índices de acierto y fallo conocidos, siguiendo la formula:

$$F_{global} = \frac{F_{acierto} \cdot \left(F_{acierto} - F_{fallo}\right)}{F_{fallo}} \tag{4.3}$$

En general se aprecia que el análisis del índice de fiabilidad conjunto no aporta mucha información adicional debido a dos causas:

a).- Los problemas que tiene interpretar este tipo de medida. La razón está en que salvo para los valores extremos es difícil dar una valoración de si una determinada palabra tiene un índice bueno o malo.

b).- Aquellas palabras que presentan muchos errores suelen tener malos índices. Esto no tendría porqué ser así, puesto que los errores de reconocimiento afectan positivamente al índice si para esos casos resultan valores próximos a cero.

Base de datos del teléfono extendido		Base de datos del proyecto IVORY		
Modo Enunciativo	***Modo Interrogativo y Exclamativo***	***Grabaciones Silenciosas***	***Grabaciones con Ruido Inducido***	***Grabaciones Ruidosas***
LPC-6	LPC-6	LPC-6	LPC-6	LPC-12
Palabra	**Palabra**	**Palabra**	**Palabra**	**Palabra**
Dos ++	Uno ++	On ++	Zero ++	Two ++
Información	Enviar	Three	On	Eight
Uno	Siete	Five	Off	On
Enviar	Mensaje	Up	Two	Last
Mensaje +	Teléfono +	Zero +	One +	Up +
...	...	...	...	...
Cancelar -	Cancelar -	Split -	Hit -	Right -
Recibir	Repetir	Yes	Last	Yes
Repetir	Establecer	Double	End	No
Cuatro	Aceptar	End	Double	Off
Aceptar --	Recibir --	Last --	No --	Next --

Tabla 4.13. Clasificación descendente por índice de fiabilidad conjunto de las palabras de diversos experimentos basados en LPC.

Base de datos del teléfono extendido		Base de datos del proyecto IVORY		
Modo Enunciativo	***Modo Interrogativo y Exclamativo***	***Grabaciones Silenciosas***	***Grabaciones con Ruido Inducido***	***Grabaciones Ruidosas***
FFT-2	FFT-5	FFT-2	FFT-2	FFT-8
Palabra	**Palabra**	**Palabra**	**Palabra**	**Palabra**
Dos ++	Mensaje ++	Three ++	Yes ++	Turn ++
Enviar	Uno	Four	Off	Zero
Siete	Tres	Jump	Six	Nine
Teléfono	Menú	Up	Right	Eight
Uno +	Cuatro +	Stop +	Stop +	Off +
...	...	...	...	...
Seis -	Seis -	Eight -	Jump -	Yes -
Ocho	Cancelar	Last	Two	On
Mensaje	Recibir	One	Ten	Up
Recibir	Aceptar	Left	Turn	Two
Aceptar --	Cero --	Right --	One --	Jump --

Tabla 4.14. Clasificación descendente por índice de fiabilidad conjunto de las palabras de diversos experimentos basados en FFT.

Base de datos del teléfono extendido				Base de datos del proyecto IVORY					
Modo Enunciativo		***Modo Interrogativo y Exclamativo***		***Grabaciones Silenciosas***		***Grabaciones con Ruido Inducido***		***Grabaciones Ruidosas***	
PLP-5 (351)		PLP-7 (371)		PLP-2 (321)		PLP-3 (331)		PLP-3 (331)	
Palabra		**Palabra**		**Palabra**		**Palabra**		**Palabra**	
Información	++	Dos	++	Last	++	Eight	++	Two	++
Siete		Mensaje		Nine		Off		Jump	
Establecer		Menú		Next		Next		GO	
Internet		Tres		Zero		Yes		One	
Repetir	+	Cero	+	Stop	+	Zero	+	Zero	+
...		...		...		...		...	
Fax	-	Cancelar	-	Off	-	Six	-	Split	-
Aceptar		Establecer		Up		Up		Six	
Mensaje		Fax		End		Jump		Off	
Tres		Recibir		Six		End		Four	
Uno	--	Aceptar	--	Go	--	Two	--	End	--

Tabla 4.15. Clasificación descendente por índice de fiabilidad conjunto de las palabras de diversos experimentos basados en PLP.

4.6. CONCLUSIONES

Las tablas Tabla 4.16 y Tabla 4.17 muestran los resultados de reconocimiento para los mejores experimentos. La primera de ellas selecciona dentro de cada conjunto de éstos (LPC, FFT y PLP), los casos individuales que aportan las mejores tasas. La segunda tabla muestra los resultados de los modelos de extracción de características con el mejor comportamiento global dentro de su grupo.

Base de datos del Teléfono extendido				Base de datos del Proyecto IVORY					
Modo Enunciativo		*Modo Interrogativo y Exclamativo*		*Grabaciones Silenciosas*		*Grabaciones con Ruido Inducido*		*Grabaciones Ruidosas*	
Modelo	**Tasa**	**Modelo**	**Tasa**	**Modelo**	**Tasa**	**Modelo**	**Tasa**	**Modelo**	**Tasa**
LPC-6	93,07%	LPC-6	94,30%	LPC-6	62,47%	LPC-6	61,57%	LPC-12	23,23%
FFT-2	92,78%	FFT-5	93,93%	FFT-2	64,90%	FFT-2	68,33%	FFT-8	26,40%
PLP-5	92,26%	PLP-7	94,00%	PLP-2	58,37%	PLP-3	57,00%	PLP-3	20,67%

Tabla 4.16. Clasificación por grupo de parametrizadores (LPC, FFT y PLP) de los mejores modelos de extracción de rasgos para cada uno de los 5 tipos de ensayo.

La primera idea que sugiere el análisis de las tasas de reconocimiento es que todas las técnicas aportan similares resultados, al menos cuando se comparan sus mejores representantes:

- El grupo de modelos LPC es mejor con la base de datos del teléfono extendido.
- El grupo de modelos FFT es mejor con la base de datos de proyecto IVORY.
- El grupo PLP aporta mejores resultados que el grupo FFT con la base de datos telefónica y mejores que el grupo LPC para los ensayos con grabaciones ruidosas.

Una explicación a este fenómeno puede ser el problema del análisis por trama fija presente en todos los esquemas de parametrización.

La gran ventaja al realizar este tipo de análisis es su perfecta compatibilidad con el paradigma establecido por los modelos ocultos de Markov. De esta forma se aprovecha toda la potencia de este esquema de reconocimiento, fundamentalmente en los aspectos relacionados la mejora en el aprendizaje cuando se dispone de mayores volúmenes de datos de entrenamiento.

Base de datos del teléfono extendido		Base de datos del proyecto IVORY		
Modo Enunciativo	***Modo Interrogativo y Exclamativo***	***Grabaciones Silenciosas***	***Grabaciones con Ruido Inducido***	***Grabaciones Ruidosas***
Experimento LPC-6				
93,07%	94,30%	62,47%	61,57%	15,90%
Experimento FFT-8				
88,93%	91,56%	62,87%	61,97%	26,40%
Experimento PLP-3				
90,89%	92,96%	57,60%	57,00%	20,67%

Tabla 4.17. Resultados de los 3 mejores modelos de extracción de parámetros en conjunto clasificados por grupo de parametrización (LPC, FFT y PLP).

La mayor desventaja tiene su origen en el conocimiento de los procesos de decodificación en los ámbitos fonético y fonológico:

a).- La caracterización de los hechos de habla no puede hacerse con periodos de tiempo fijo.

b).- El estudio de los sonidos presentes en una realización de habla debe ser claramente relacional, apoyando las conclusiones que se puedan obtener para determinados fragmentos con los resultados de las otras partes. En este sentido es importante notar, que no existe un tratamiento diferenciado entre vocales y consonantes, cuando las primeras son más importantes y sirven de guía en la interpretación de éstas otras.

Otra conclusión que puede extraer del análisis de los resultados presentes en la

Tabla 4.17 es que ningún modelo de parametrización ofrece el mejor comportamiento en todas las situaciones, tanto si lo comparamos con otros de su grupo o en conjunto.

Los mejores esquemas de extracción de parámetros son aquellos en los que mejor se relaciona la información presente en las diferentes bandas del espectrograma.

Esto es especialmente claro para los experimentos LPC o para los casos que emplean FFT y coeficientes cepstrales, donde la información presente en las citadas bandas de frecuencia se combina para obtener medidas de la distribución de la información.

Los experimentos basados en PLP emplean, en esencia, el mismo principio que los casos de coeficientes cepstrales con el elemento añadido de realizar un análisis de predicción lineal una vez que ha reducido la información presente en el espectro a un número finito de bandas de información. Este último paso permite obtener una descripción compacta, aunque muy difícil de interpretar.

Por otra parte, es destacable el hecho de que los mejores extractores de rasgos, sean aquellos que incluyen características dinámicas. Esto no indica que las características dinámicas, así calculadas, sean más importantes que las estáticas, simplemente que resultan ser un muy buen complemento de éstas.

No obstante, su concurso se muestra imprescindible y en todo caso se podría hacer la misma objeción que antes, es decir, el cálculo de dichos coeficientes tomando periodos de tiempo prefijados, por mucho que se busque un valor óptimo, no parece ser la mejor solución que se pueda alcanzar para abordar la tarea del análisis fonético de la señal de voz.

Todos los sistemas de obtención de características ofrecen resultados mucho más pobres en el caso de las grabaciones ruidosas que para los otros dos conjuntos de test. En este sentido se puede comprobar la falta de robustez de todos ellos. También, hay que tener en cuenta, que los errores de borrado, muy frecuentes aquí, no aparecen recogidos en estos datos.

Comparando todo tipo de grabaciones realizadas en un ambiente silencioso, los resultados con la base de datos del teléfono extendido son mucho mejores en todos los casos:

- Existe mucha más variabilidad debido a que los locutores producen realizaciones en una lengua distinta de la materna.
- Hay más datos de entrenamiento disponibles para la base de datos del teléfono extendido, a pesar de que el número de locutores para entrenar es menor. Sin embargo, aún equilibrando este aspecto la conclusión no varía, simplemente se reduce la diferencia entre ambos grupos.

Los resultados para los ensayos con entonación enunciativa y para los de entonación exclamativa/interrogativa indican, que cuando se entrena con ambos casos, la calidad de la respuesta es prácticamente idéntica. Lo mismo sucede con los resultados para las pruebas con ambientes silencioso y de ruido inducido.

En los resultados de reconocimiento parciales se aprecia, por regla general, que la distribución equilibrada de locutores (masculinos y femeninos), así como, la de realizaciones de habla de los diferentes grupos (según el ambiente de grabación o el tipo de entonación), favorece un mejor comportamiento de los sistemas de reconocimiento.

En todos los casos, sin excepción, se obtienen mejores resultados cuantos más datos hay para entrenar. Esto sucede incluso en aquellos sistemas en los que la aportación de nuevos ejemplos hace descender mucho sus tasas de reconocimiento para el test de los datos del entrenamiento (LPC-1). Además en nuestros experimentos esta realidad se ve amplificada por el hecho de que las bases de datos empleadas son de tamaño pequeño.

Un aspecto muy relacionado con el punto anterior es el hecho de que en todos los casos se verifica, que el test sobre los datos del entrenamiento ofrece peores resultados, según se incorporan nuevos ejemplos. En principio esto no es grave si la pérdida es pequeña, aunque da una idea clara de la capacidad de discriminación de éstos.

Los índices de fiabilidad son muy buenos para los entrenamientos pero poco útiles

en el resto de casos:

a).- Las diferencias entre los índices de acierto y fallos son muy pequeñas, incluso negativas para algunos casos con realizaciones ruidosas.

b).- La arquitectura empleada (extractor de rasgos+ cuantificador vectorial+ *parser* HMM), no permite extraer ninguna medida que pueda ayudar a determinar si la palabra reconocida es correcta. Este es un problema, no sólo de nuestro sistema, sino de los reconocedores de voz en general.

Esto implica, además, que no existe ninguna forma eficaz de obtener segundas opiniones y menos aún, en las primeras etapas del reconocimiento. Hay que tener en cuenta que en nuestro caso no hay ninguna gramática porque cualquier palabra puede aparecer en cualquier momento, lo que vienen a significar, que si el análisis fonético es erróneo, no hay forma de corregir el fallo en fases posteriores usando conocimiento del contexto. En la práctica, la única posibilidad se reduce a tomar la palabra con la segunda mejor probabilidad otorgada por los modelos de Markov.

5. APLICACIÓN DE INFORMACIÓN ACÚSTICA A LA ETAPA DE EXTRACCIÓN DE PARÁMETROS

5.1. INTRODUCCIÓN

Los esfuerzos por incorporar de manera directa conocimiento fonético-acústico a la fase de extracción de características y por lo general al proceso de reconocimiento automático del habla, aparecen caracterizados por una serie de factores que conviene analizar. A lo largo del presente capítulo estudiaremos los aspectos clave que determinan este punto.

En esencia, son cuatro las variables que determinan el éxito o el fracaso de este tipo de operaciones:

1. Posibilidad de aportar información relevante o discriminante.
2. Fiabilidad en el proceso de incorporación del nuevo conocimiento.
3. Correlación e incorrelación con otras fuentes de conocimiento empleadas durante la extracción espectral.
4. Compatibilidad con el resto de las etapas del sistema reconocedor.

El primer aspecto, el relativo a poder aportar información útil al sistema de reconocimiento, viene determinado por el grado de comprensión de los procesos de producción y percepción del habla.

Si bien la aplicación del saber, proporcionado por disciplinas tales como la Fonética y la Fonología, ha sido insuficiente en el pasado, lo cierto es que, por lo general, existe una clara falta de generalidad en muchas de sus conclusiones. Además, en la mayoría de los casos, se desconoce el grado de invarianza que pueden alcanzar muchos de los criterios de decisión y análisis empleados.

La fiabilidad en la incorporación de nuevas fuentes de conocimiento aparece claramente relacionada con la posibilidad real de disponer de herramientas de análisis útiles y robustas. Un buen ejemplo de este apartado lo constituyen los métodos de búsqueda de formantes. Los errores propios de los algoritmos de detección de formantes

neutralizan la mejora introducida por dichos métodos [HOL97], [GAR98].

Además, está el problema de la relevancia de la información en determinados periodos de tiempo, como sucede por ejemplo con los sonidos sordos para el tipo de métodos citado anteriormente. Por consiguiente, tan importante como poder disponer de nuevas fuentes de conocimiento, es poder conocer los límites de aplicación de éstas.

Por su parte, el problema de la correlación/incorrelación entre las características contenidas en un problema está presente en cualquier problema de reconocimiento. En el caso de la voz, muchas de las bondades de determinados modelos de parametrización, como por ejemplo los coeficientes cepstrales obtenidos a partir de la transformada de Fourier, tienen su explicación en la escasa correlación entre coeficientes [FUR89].

Una fuerte correlación o redundancia entre los rasgos extraídos puede ser el origen de reducciones drásticas en las capacidades de discriminación de un sistema de reconocimiento. Si además intentamos concatenar informaciones procedentes de diversos orígenes este problema puede ser de la máxima importancia.

Por último, un aspecto clave que va a determinar las prestaciones de un esquema de extracción de rasgos viene definido por el grado de compatibilidad con el resto de las etapas del proceso. En el caso concreto del reconocimiento automático de discurso, la parte susceptible de causar más conflictos con la fase de extracción de características es la encargada de realizar el agrupamiento y normalización estadística de los datos de la entrada. En nuestros experimentos, esta fase está conformada por el proceso de cuantificación vectorial, aunque en general, encontraremos un conjunto de mezclas de funciones de densidad de probabilidad.

La compatibilidad entre las distintas tareas, en lo que atañe a la etapa de obtención de vectores acústicos, vendrá determinada por la idoneidad en la unión de canales de información radicalmente distintos en cuanto a sus propiedades:

1. Nivel de abstracción.

2. Rango numérico.

3. Sensibilidad a determinados fenómenos acústico-fonéticos.
4. Estabilidad y velocidad de variación a lo largo del tiempo (tramas consecutivas).

Por lo general, deberán agruparse en un mismo subconjunto, todas aquellas características o rasgos que compartan la totalidad o parte de las propiedades mencionadas. Así, por ejemplo, no parece muy sensato juntar características de tipo general, como son aquellas que permiten decidir entre bloques de voz sonoros y sordos, con otras particulares, como las que pueden servir para determinar el punto de articulación en fragmentos de voz correspondientes a un sonido oclusivo.

En las siguientes secciones vamos a describir algunos aspectos que vamos a analizar de entre los citados anteriormente. En concreto serán:

a).- Reducción dimensional.

b).- Cuantificación vectorial jerárquica.

c).- Aplicación del tono fundamental.

5.2. REDUCCIÓN DIMENSIONAL

Habitualmente, la motivación existente para aplicar métodos de reducción de dimensión en los vectores de características de los sistemas de reconocimiento, viene dada por el hecho de que la presencia de componentes fuertemente correladas con el resto de los elementos del vector, va a producir pérdidas notables en las capacidades discriminativas del sistema de reconocimiento.

Algunos de los métodos empleados comúnmente hacen uso del cálculo de autovalores con el objetivo de poder determinar aquellas componentes del vector de rasgos que son más independientes del resto, o dicho de otra manera, que no pueden deducirse de las demás; así como aquellas que por el contrario pueden obtenerse de forma fácil a partir de otras. Algunos ejemplos de técnicas frecuentemente aplicadas son: el análisis de componentes principales o PCA *(Principal Component Analysis)* y el análisis discriminante lineal o LDA (*Linear Discriminat Analysis*) [DIA99].

En la práctica, estos procedimientos resultan muy útiles puesto que permiten realizar todo el proceso partiendo únicamente de los datos, es decir, partiendo de conjuntos reales de vectores de parámetros. Además, no son exclusivos del ámbito del reconocimiento del habla.

En nuestro caso, sin embargo, la motivación es diferente. El planteamiento para afrontar una posible reducción, o cambio en la composición de un vector de rasgos fonéticos, vendrá provocado por un análisis acerca del tipo de información fonético-acústica, que porta cada componente de dicho vector. Es decir, si un determinado rasgo no tiene un significado fonético deducible, deberá ser un claro candidato a desaparecer del sistema.

El mayor inconveniente de este enfoque es que muchos de los sistemas de parametrización espectral no permiten realizar este análisis. Algunos ejemplos de esto último lo encontramos en los modelos de extracción de parámetros basados en coeficientes LPC y LPC-cepstrum, así como los basados en PLP cuando contienen la

parte del análisis de predicción lineal. Para todos estos casos la tarea de intentar asociar un significado a cada componente del vector de rasgos resulta ser muy complicada, por no decir imposible. Esto implicará que la mejora de dichos sistemas no podrá realizarse por medio de un mejor conocimiento del problema.

No obstante, sí podemos aplicar estas ideas a buena parte de los parametrizadores que se consideran determinantes del estado de la cuestión y para los que obtuvimos algunos de los mejores resultados de reconocimiento.

Si analizamos el significado que tienen los métodos basados en la transformada de Fourier y coeficientes cepstrales, por ejemplo, comprobamos cómo cada coeficiente cepstral representa una combinación lineal de la información presente en cada banda de frecuencias. Por tanto, el significado acústico, que tiene la transformada discreta del coseno es el de relacionar de diferentes formas el contenido presente en cada grupo de frecuencias. El análisis de estos procesos de combinación lineal sobre el contenido frecuencial de la señal permite valorar la importancia que cada cepstrum tiene en la descripción total proporcionada por el vector de parámetros.

En la Figura 5.1 aparecen representadas las curvas que muestran la ponderación que se aplica a cada banda de energía (20 en total) para cada uno de los 10 coeficientes cepstrales empleados en nuestros experimentos.

Así el primer coeficiente cepstral recogerá la proporción entre la energía contenida en las frecuencias bajas (< 1000 Hz) y las regiones medias y altas. El segundo coeficiente contendrá la relación existente entre el contenido espectral medio (región contenida entre los 500 y los 2500 Hz) y el resto de las frecuencias. Por su parte el tercer coeficiente separa las regiones situadas entre los 500-1000 Hz y 3250-5500 Hz del resto y así sucesivamente.

Se puede comprobar cómo los primeros coeficientes presentan relaciones entre las distribuciones de energía frecuencial más claras de interpretar. Los últimos coeficientes, si bien, marcan diferencias más finas en cuanto a los contenidos espectrales, son mucho más difíciles de poder caracterizar.

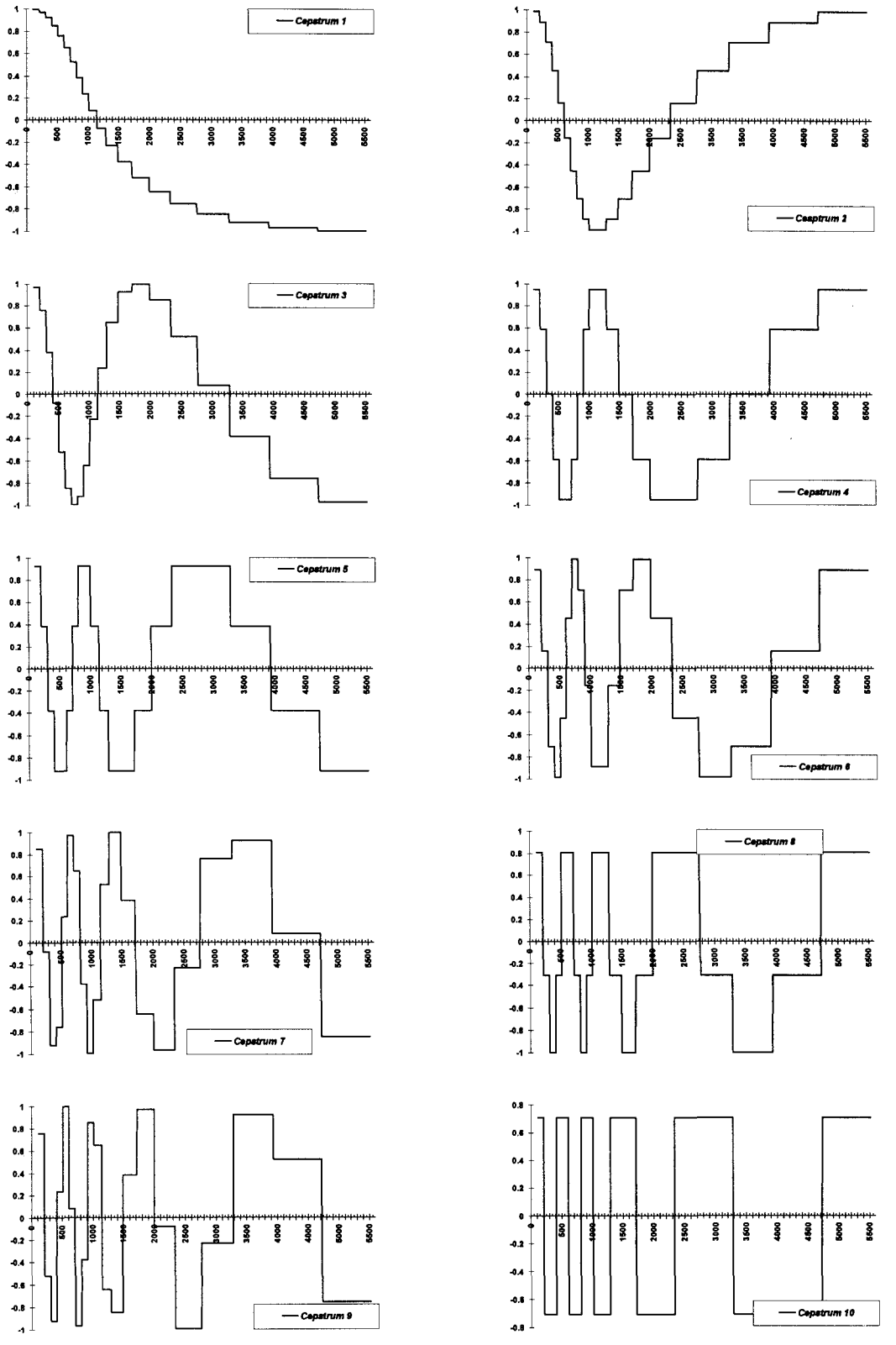

Figura 5.1. Respuesta en frecuencia, que sobre las 20 bandas de frecuencia de los experimentos descritos tienen los 10 primeros coeficientes cepstrales.

La gran ventaja, sin embargo, es que todos están poco correlados entre sí. Esto significa que si incorporamos elementos (coeficientes) poco representativos como aparentan ser los últimos coeficientes cepstrales, las capacidades de reconocimiento del sistema no debieran resentirse mucho.

Como forma de corroborar en parte nuestras ideas, planteamos más adelante algunos experimentos de reconocimiento en los cuales se ha reducido de forma progresiva el número de coeficientes cepstrales.

5.3. CUANTIFICACIÓN VECTORIAL JERÁRQUICA

La motivación para realizar la cuantificación vectorial jerárquica hace referencia a la necesidad de incorporar informaciones o conocimientos, que presentan distintos niveles de abstracción y/o generalidad durante la fase de construcción de vectores acústicos. Un rasgo espectral que puede ser útil en determinadas circunstancias puede no serlo en otras, o cuando menos, estar supeditado a la presencia de determinados valores de otro u otros rasgos. Si esa información aparece siempre en el vector de características, en algunos momentos enmascarará el contenido de otras propiedades que para esa situación son más relevantes.

Una posibilidad para verificar la mayor o menor importancia de determinadas componentes del vector de características, es estudiar la distribución de clases como consecuencia del proceso de cuantificación vectorial. Durante el proceso de creación de los libros de códigos, cada componente espectral se pondera por el valor de su varianza. Con esta corrección se evita que la cuantificación vectorial esté dominada por la/s componente/s con mayor varianza, a la hora de determinar las medidas de distorsión o distancia entre vectores.

Tomando un sencillo ejemplo, como el que aparece en la Figura 5.2, encontramos que la serie completa de vectores *(x, y)* se puede agrupar perfectamente en 4 clases (Figura 5.3). La Tabla 5.1 muestra que la varianza de la componente *y* es mucho mayor, pero visualmente se comprueba que es ésta componente la que mejor permite asociar cada vector *(x, y)* con su respectiva clase.

Vamos a establecer una medida de separación entre clases para todas las componentes del vector de parámetros. Este criterio se basa en la medida discriminante de Fisher [NIC97]:

$$Sep(i,j,x) = \frac{\left(\mu_{i(x)} - \mu_{j(x)}\right)^2}{\sigma^2_{i(x)} + \sigma^2_{j(x)}} \tag{5.1}$$

siendo i e j dos clases de vectores y x la dimensión n-ésima del vector de parámetros.

Por el resto de los resultados presentes en la Tabla 5.1 se puede apreciar como la componente x solamente es mejor para separar a la clase D de las clases B y C, en el resto de clases la otra componente es claramente superior.

Por tanto, podemos emplear la medida de separación media entre clases como criterio para determinar cuales son las componentes más importantes e influyentes a la hora de asociar los vectores de parámetros a sus respectivas clases. La separación media entre clases puede ser empleada siguiendo una filosofía similar a la descrita en el apartado anterior (reducción de dimensionalidad), o puede servir como herramienta de análisis con objeto de poder establecer un procedimiento de cuantificación vectorial jerárquica o por niveles.

	Eje X	**Eje Y**
Varianza	*0.68769611*	*9.2791131*
Centroide Clase A	0.953704	9.55400046
Centroide Clase B	2.007722	5.90641949
Centroide Clase C	2.448118	1.96256207
Centroide Clase D	0.913481	3.85328649
Varianza Clase A	*0.42843506*	*1.27820494*
Varianza Clase B	*0.38337502*	*1.45270796*
Varianza Clase C	*0.08718165*	*1.25616931*
Varianza Clase D	*0.0816553*	*1.39612896*
Separación (A,B)	1.36848996	4.87194115
Separación (A,C)	4.33126616	22.739316
Separación (A,D)	0.00317177	12.1518632
Separación (B,C)	0.4121685	5.74186637
Separación (B,D)	2.57480711	1.47967582
Separación (C,D)	13.9490247	1.34782685
Separación Media Clases	*3.77315471*	*8.0554149*

Tabla 5.1. Resultados para la distribución de vectores bidimensionales de las 4 clases de la Figura 5.2.

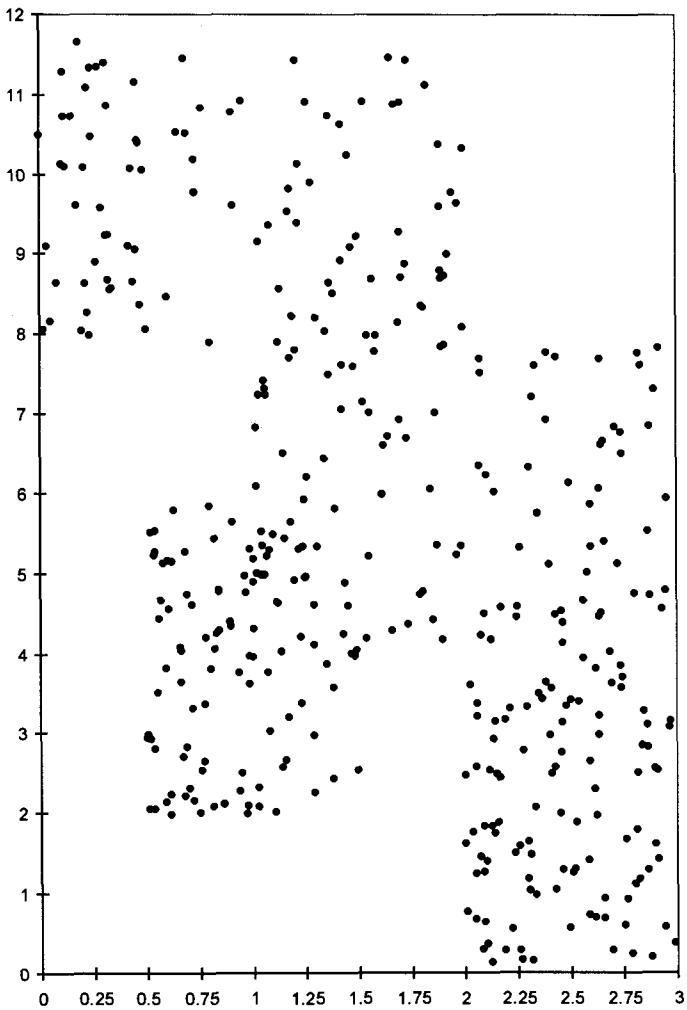

Figura 5.2. Ejemplo de una distribución de vectores bidimensionales.

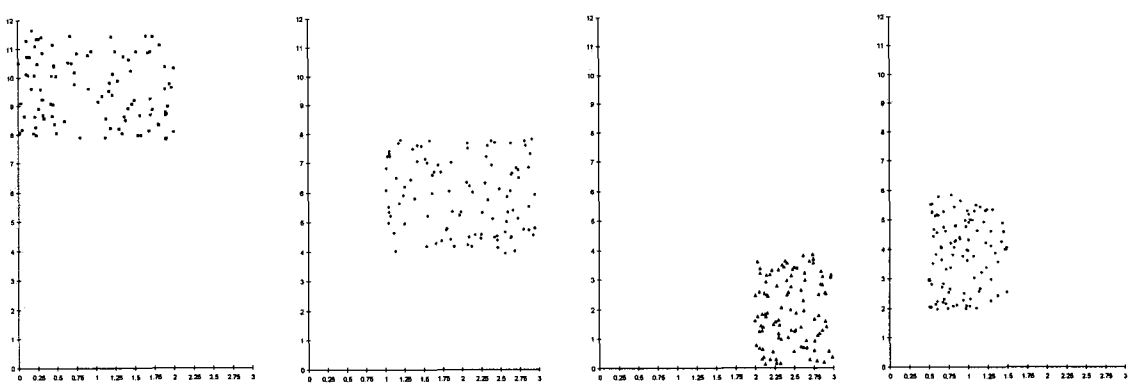

Figura 5.3. Distribución de valores *(x, y)* para las 4 clases (A, B, C, D) definidas en la Figura 5.anterior.

El método de cuantificación vectorial por niveles permite durante el proceso de cuantificación vectorial separar por grupos o niveles aquellas componentes del vector de parámetros que presenten un nivel de generalidad similar y cuyas dependencias con otras componentes deba ser mantenida. En definitiva se trata de conseguir que aquellas componentes que describan propiedades generales (ej. sordo, sonoro) no convivan con otras que aun siendo útiles son más particulares (ej. punto de articulación en oclusivas sordas).

El método, que describimos (Figura 5.4), pasa por emplear un primer grupo de rasgos para obtener un conjunto pequeño de clases (nivel 1). En un segundo nivel, se utilizarían otros rasgos distintos para obtener una serie de subclases a partir de las primeras, pero utilizando solamente los vectores de entrada que estuvieran asociados a esas clases del primer nivel. Así sucesivamente hasta finalizar el último nivel.

El procedimiento de cálculo es idéntico al procedimiento habitual con la salvedad de que no se emplean todas las componentes del vector de parámetros. Las características esenciales de dicho método son:

a).- En cada nivel se emplea un subconjunto de rasgos/componentes distinto cada vez. Cada rasgo estará asociado a un único nivel de cuantificación.

b).- Una vez finalizado un determinado nivel y generadas un total de N_{nivel} clases, el proceso con el siguiente nivel equivaldría a tener N_{nivel}-problemas de cuantificación independientes, con su parte correspondiente de los vectores de entrada.

c).- Las clases que se calculan en un nivel van completándose en los siguientes, es decir, añadiendo nuevas componentes a sus vectores centroide (exactamente las correspondientes a dichos niveles).

d).- Con objeto de generar nuevas clases, se produce un fenómeno de división en dos. En este último caso, las componentes del vector asociadas a niveles superiores en la jerarquía o inferiores en número de nivel, se mantienen constantes durante la división, es decir, la separación en clases calculada para niveles anteriores no varía aunque el número de clases vaya creciendo.

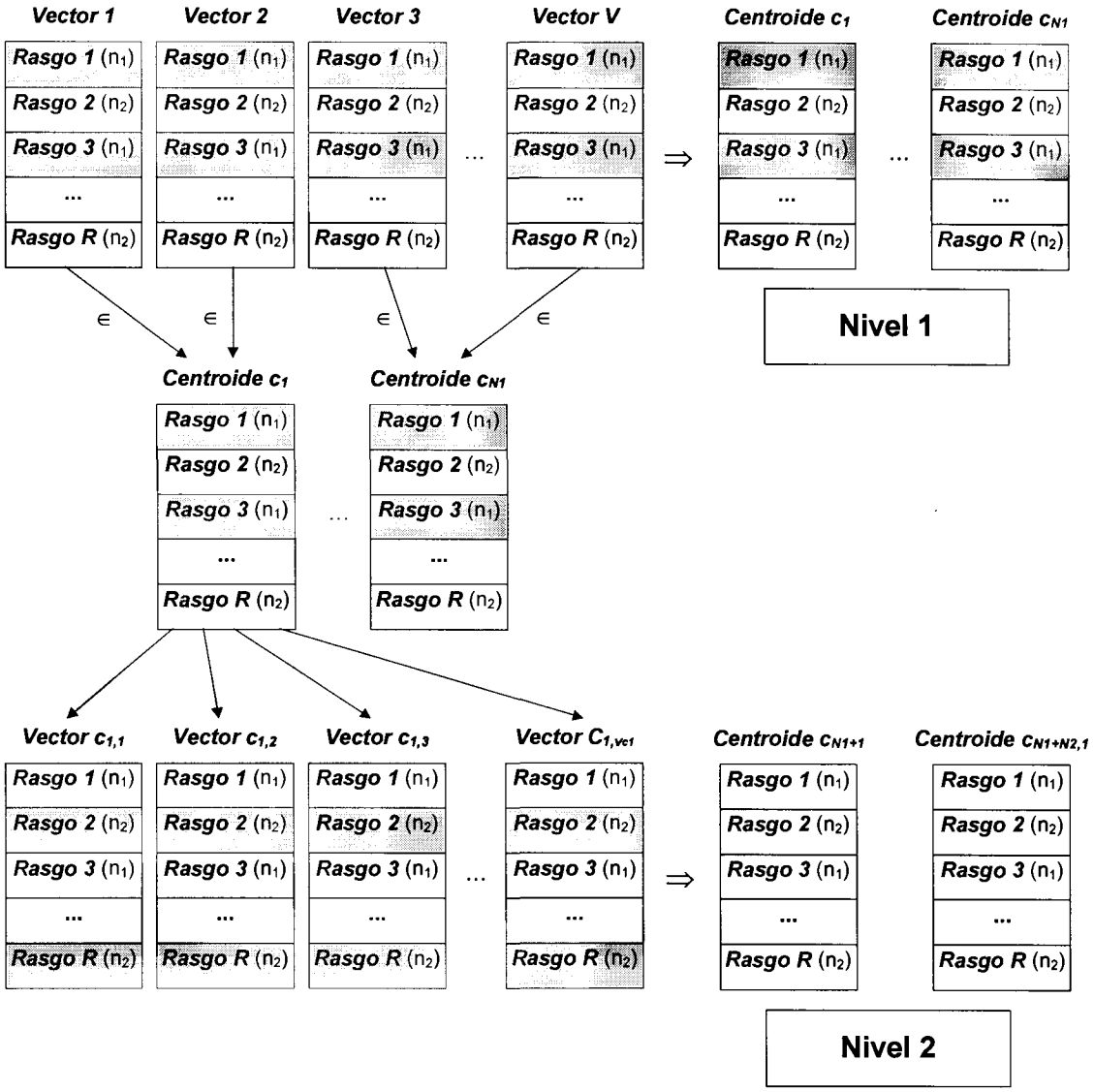

Figura 5.4. Ejemplo de proceso de cuantificación vectorial en 2 niveles. Durante el primer nivel se realiza la obtención de N1 centroides a partir del conjunto de rasgos asociados a este nivel (n1). Durante el segundo nivel se crean nuevos centros derivados de los anteriores, hasta alcanzar el valor N2, a partir de las componentes n2 (manteniendo constantes los rasgos n1).

Al igual que para el apartado anterior, analizaremos la aplicación de dicho método a diferentes modelos de extracción de parámetros.

5.4. APLICACIÓN DEL TONO FUNDAMENTAL

Presentamos en esta sección un algoritmo de cálculo del tono fundamental, basado en técnicas de programación dinámica, que se ha utilizado como parámetro adicional en algunos de los experimentos de reconocimiento presentados en esta sección.

Su objetivo es ilustrar cómo la aplicación al proceso de reconocimiento de fuentes de conocimiento relevantes, aunque heterogéneas, puede servir para mejorar sus prestaciones. Por lo general, los sistemas de reconocimiento no incluyen información del tono fundamental [KOM97].

El algoritmo, basado en el análisis de los máximos de la traza de voz, consta de las siguientes fases:

1. Cálculo de los máximos/picos presentes en la señal.
2. Cálculo de la función de coste del conjunto de máximos anteriores.
3. Construcción de los caminos de mínimo coste.

Los picos pertenecientes a caminos de coste mínimo se tomarán como los orígenes de cada periodo de la señal de voz. El valor del tono fundamental de una trama se calculará a partir de las separaciones temporales existentes entre los picos presentes en dicha trama y los máximos adyacentes a éstos en el camino elegido.

Un pico/máximo solamente puede formar parte de un camino. Sin embargo, una trama a efectos de reconocimiento puede contener varios picos de uno o varios caminos. En este caso el valor del tono fundamental vendrá dado por una combinación entre varios valores, próximos por otra parte.

5.4.1. Cálculo de los máximo/picos de la señal de voz

Obtiene todos los máximos positivos de la señal de voz.

Algoritmo:

```
numeroPicosTotales= 0
<Para> i= 1 <Hasta> i= numeroMuestrasVozTot
        <Si> muestra(i+1)> 0
                <Si> muestra(i+1)> muestra(i) y muestra(i+1)> muestra(i+2)
                        <Almacenar pico>
                        <Incrementar> numeroPicosTotales
                        <Almacenar pico>
```

5.4.2. Cálculo de la función de coste del conjunto de máximos de la señal de voz.

Se calcula para todos los picos existentes, la correspondiente función de coste.

Algoritmo:

```
<Para> k= N_PICOS_CONTEXTO <Hasta> k= (numeroPicosTotales- N_PICOS_CONTEXTO)
        coste= CalcularCostePico(k, mejorAnterior, mejorPosterior)
        tonoFundamental= mejorTonoFundamental;
        <Si> coste< UMBRAL_COSTE_TONO_FUNDAMENTAL
                <Almacenar información> pico p_k
                        k
                        coste
                        mejorAnterior;
                        mejorPosterior;
                        tonoFundamental
```

5.4.2.1. Calcular la función de coste para un máximo de la señal de voz

Esta función calcula el coste asociado a un determinado máximo (*picoActual*) cuando se considera que éste es el comienzo de un nuevo periodo de la señal de voz. Realiza la búsqueda tomando dos conjuntos de máximos próximos a *picoActual*. Obtiene como salida el valor de la función de coste del máximo, así como, los otros 2 picos con los que se obtienen los mejores resultados (*picoAnterior, picoPosterior*).

Algoritmo:

1. Inicio:

```
valorcosteMinimo= ∞
mejorAnterior= -1
mejorPosterior= -1
```

2. Bucle principal:

```
<Para> i= (picoActual-1) <Hasta> (picoActual-N_PICOS_CONTEXTO)
        <Para> j= (picoActual+1) <Hasta> (picoActual+N_PICOS_CONTEXTO)
                coste= Calcular funcion de coste para 3 máximos(i, pico, j)
                <Si> coste< (costeMinimo*FACTOR_MEJORA)
                        valorCosteMinimo= coste;
                        mejorAnterior= i;
                        mejorPosterior= j;
                        <Si> coste< UMBRAL_SOLUCION_VALIDA
                                <Finalizar proceso>

picoAnterior= mejorAnterior;
picoPosterior= mejorPosterior;
```

El par de máximos (uno por la izquierda y otro por la derecha) obtenidos de forma adecuada, con objeto de proporcionar al *picoActual* un valor de coste pequeño, debe estar lo más próximo posible a éste. La idea es evitar calcular valores del tono fundamental que sean múltiplos de su valor real. De ahí que se establezca un factor de mejora para picos más alejados en el tiempo.

5.4.2.2. Calcular la función de coste para 3 máximos(p_i, p_j, p_k)

Dados 3 máximos tomados de la señal de voz en el dominio del tiempo p_i, p_j y p_k denotaremos como mp_i, mp_j y mp_k a sus respectivos números de muestra dentro de la señal de voz a la que pertenecen. Por su parte, vp_i, vp_j y vp_k serán los valores que alcanzan dichos máximos (Figura 5.5).

Definimos entonces una función de coste que determina la bondad en la consideración de que el inicio de 3 periodos consecutivos de la señal dan comienzo en esas posiciones. Su definición es:

$$C(p_i,p_j,p_k)=\begin{cases} C_T(p_i,p_j,p_k) & Si \quad vp_i>0,\ vp_j>0,\ vp_k>0 \\ & \quad J(p_i,p_{i+1},p_{i+2})<Max.jitter \\ & \quad Min.d<d(p_i,p_j)<Max.d \\ & \quad Min.d<d(p_j,p_k)<Max.d \\ \infty & resto \end{cases} \tag{5.2}$$

Por tanto, la función de coste solamente tendrá validez cuando:

a).- Los máximos sean positivos.

b).- La separación dos a dos o *jitter* (variación en el tono fundamental a lo largo del tiempo) esté por debajo de un umbral

c).- La separación entre dos picos o máximos esté dentro de los umbrales, que limitan los valores permisibles que el periodo puede alcanzar.

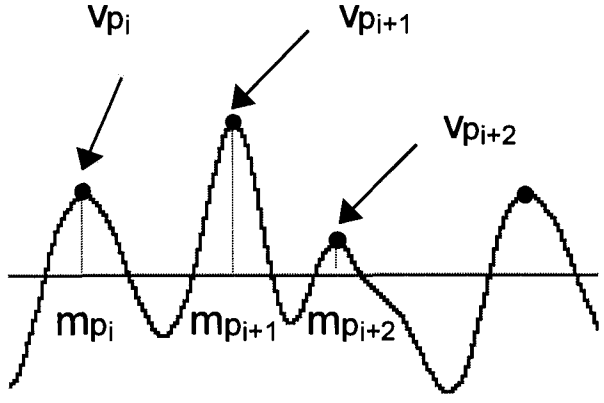

Figura 5.5. Detalle del procedimiento de cálculo del tono fundamental.

La función de coste valora a su vez dos aspectos:

$$C_T(p_i,p_j,p_k)=C_1(p_i,p_j,p_k)\cdot C_2(p_i,p_j,p_k) \tag{5.3}$$

El primer término $C_1(p_i, p_j, p_k)$ expresará las relaciones existentes entre los 3 máximos en cuanto a su separación en el tiempo:

$$C_1(p_i,p_j,p_k)=\begin{cases}1.0 & Si\ d(p_i,p_j)=d(p_j,p_k)\\ P(p_i,p_j,p_k)\cdot J(p_i,p_j,p_k) & resto\end{cases} \quad (5.4)$$

siendo:

$$d(p_i,p_j)=\begin{cases}mp_i-mp_j & i\geq j\\ mp_j-mp_i & resto\end{cases} \quad (5.5)$$

$$J(p_i,p_j,p_k)=\left|d(p_i,p_j)-d(p_j,p_k)\right| \quad (5.6)$$

$$P(p_i,p_j,p_k)=\begin{cases}\dfrac{d(p_i,p_j)}{d(p_j,p_k)} & Si\ d(p_i,p_j)\geq d(p_j,p_k)\\ \dfrac{d(p_j,p_k)}{d(p_i,p_j)} & resto\end{cases} \quad (5.7)$$

El segundo término $C_2(p_i, p_j, p_k)$ pondera la separación entre los valores de las muestras de los 3 máximos analizados:

$$C_2(p_i,p_j,p_k)=\frac{dv(p_i,p_j)}{vp_i\cdot vp_j}+\frac{dv(p_j,p_k)}{vp_j\cdot vp_k} \quad (5.8)$$

siendo *dv*:

$$dv(p_i,p_j)=\left|vp_i-vp_j\right| \quad (5.9)$$

5.4.2.3. Calcular el tono fundamental para 3 máximos(p_i, p_j, p_k)

Conocidos 3 máximos tomados de la señal de voz en el dominio del tiempo p_i, p_j y p_k, el valor del tono fundamental asociado vendrá dado por la siguiente expresión:

$$T.F.=\frac{d(p_i,p_j)+d(p_j,p_k)}{2} \quad (5.10)$$

El valor obtenido estará referido a un número de muestras de separación. Conocida la frecuencia de muestreo puede calcularse el valor del tono fundamental en Hz si se desea.

5.4.3. Construcción de caminos de mínimo coste

Partiendo de cada pico existente se construye el camino de mínimo coste. Un pico solamente puede pertenecer a un solo camino y debe poseer un coste menor que el marcado por un umbral. Un pico podrá añadirse a un camino cuando el valor de *mejorPosterior* del último pico de un camino en construcción coincida con el valor *mejorAnterior* de este otro *mejorPosterior*. Se desecharán todos los caminos que al menos no contengan 2 valores.

Algoritmo:

<Para> j=N_PICOS_CONTEXTO <Hasta> numeroPicosTotales- N_PICOS_CONTEXTO
    <Recuperar información> pico p_j
    <Si> p_j no forma parte de ningún camino
        <Si> coste p_j < UMBRAL_COSTE_TONO_FUNDAMENTAL
            longitudCamino de p_j = 0
            siguientePico= j
            <Hacer>
                picoActual= siguientePico
                siguientePico= mejorPosterior al p_j
                <Recuperar información> pico $p_{siguientePico}$
                <Si> valorCostePico< UMBRAL_COSTE_TONO_FUNDAMENTAL
                    <Incrementar> longitudCamino de p_j
                    siguientePico= mejorPosterior al $p_{siguientePico}$
            <Mientras> picoActual= mejorAnterior al $p_{siguientePico}$

5.5. ESTUDIO DE SU APLICACIÓN A LOS EXPERIMENTOS DE RECONOCIMIENTO DEL HABLA

5.5.1. Descripción de los nuevos experimentos

Es este apartado mostramos el conjunto de 16 nuevos experimentos de reconocimiento, donde incorporamos los aspectos que se señalaron anteriormente.

Todos los casos que presentamos en este apartado tienen como punto de partida al experimento FFT-8. En esencia los nuevos casos se van a agrupar en tres grandes grupos:

1. Experimentos donde se reduce el número de coeficientes cepstrales empleado (Tabla 5.2).
2. Experimentos con diferente número de coeficientes y usando cuantificación vectorial con varios niveles (Tabla 5.3).
3. Experimentos donde se incluya la información del tono fundamental. El proceso de cuantificación se realiza con uno o más niveles (Tabla 5.4).

Para los puntos 2 y 3 se establecen diferentes particiones del número de clases asociado con cada nivel de la cuantificación.

Código Parametrización	Descripción	Número y Tipo de Parámetros
FFT-8	FFT-Cepstrum	10 coeficientes + 10 coeficientes delta (5 tramas).
FFT-$8_{(1)}$		6 coeficientes + 6 coeficientes delta (5 tramas).
FFT-$8_{(2)}$		4 coeficientes + 4 coeficientes delta (5 tramas).
FFT-$8_{(3)}$		2 coeficientes + 2 coeficientes delta (5 tramas).

Tabla 5.2. Conjunto de experimentos basados en el parametrizador FFT-8, empleando diferente número de coeficientes cepstrales. El primer caso corresponde con el caso original.

Código Parametrización	Descripción	Número y Tipo de Parámetros
FFT-$8_{(4)}$	FFT-Cepstrum	10 coeficientes + 10 coeficientes delta (5 tramas). Cuantificación en 2 niveles: 1ª nivel [6 coeficientes + 6 coeficientes delta (5 tramas)], (16 índices). 2ª nivel [4 coeficientes + 4 coeficientes delta (5 tramas)], (240 índices).
FFT-$8_{(5)}$		10 coeficientes + 10 coeficientes delta (5 tramas). Cuantificación en 3 niveles: 1ª nivel [4 coeficientes + 4 coeficientes delta (5 tramas)], (16 índices). 2ª nivel [2 coeficientes + 2 coeficientes delta (5 tramas)], (48 índices). 3ª nivel [4 coeficientes + 4 coeficientes delta (5 tramas)], (192 índices).
FFT-$8_{(6)}$		10 coeficientes + 10 coeficientes delta (5 tramas). Cuantificación en 4 niveles: 1ª nivel [2 coeficientes + 4 coeficientes delta (5 tramas)], (16 índices). 2ª nivel [2 coeficientes + 2 coeficientes delta (5 tramas)], (48 índices). 3ª nivel [2 coeficientes + 2 coeficientes delta (5 tramas)], (64 índices). 4ª nivel [4 coeficientes + 4 coeficientes delta (5 tramas)], (128 índices).
FFT-$8_{(7)}$		4 coeficientes + 4 coeficientes delta (5 tramas). Cuantificación en 2 niveles: 1ª nivel [2 coeficientes + 2 coeficientes delta (5 tramas)], (64 índices). 2ª nivel [2 coeficientes + 2 coeficientes delta (5 tramas)], (192 índices).
FFT-$8_{(8)}$		4 coeficientes + 4 coeficientes delta (5 tramas). Cuantificación en 3 niveles: 1ª nivel [1 coeficiente + 1 coeficiente delta (5 tramas)], (16 índices). 2ª nivel [1 coeficiente + 1 coeficiente delta (5 tramas)], (48 índices). 3ª nivel [2 coeficientes + 2 coeficientes delta (5 tramas)], (192 índices).
FFT-$8_{(9)}$		4 coeficientes + 4 coeficientes delta (5 tramas). Cuantificación en 4 niveles: 1ª nivel [1 coeficiente + 1 coeficiente delta (5 tramas)], (16 índices). 2ª nivel [1 coeficiente + 1 coeficiente delta (5 tramas)], (48 índices). 3ª nivel [1 coeficiente + 1 coeficiente delta (5 tramas)], (64 índices). 4ª nivel [1 coeficiente + 1 coeficiente delta (5 tramas)], (128 índices).

Tabla 5.3. Conjunto de experimentos basados en el parametrizador FFT-8, realizando el proceso de cuantificación vectorial por niveles.

Código Parametrización	Descripción	Número y Tipo de Parámetros
FFT-$8_{(10)}$	Tono fundamental + FFT-Cepstrum	Tono fundamental + 4 coeficientes + 4 coeficientes delta (5 tramas).
FFT-$8_{(11)}$		Tono fundamental + 4 coeficientes + 4 coeficientes delta (5 tramas). Cuantificación en 2 niveles: 1ª nivel [Tono fundamental], (2 índices). 2ª nivel [4 coeficientes + 4 coeficientes delta (5 tramas)], (254 índices).
FFT-$8_{(12)}$		Tono fundamental + 4 coeficientes + 4 coeficientes delta (5 tramas). Cuantificación en 2 niveles: 1ª nivel [Tono fundamental], (4 índices). 2ª nivel [4 coeficientes + 4 coeficientes delta (5 tramas)], (252 índices).
FFT-$8_{(13)}$		Tono fundamental + 4 coeficientes + 4 coeficientes delta (5 tramas). Cuantificación en 2 niveles: 1ª nivel [Tono fundamental], (8 índices). 2ª nivel [4 coeficientes + 4 coeficientes delta (5 tramas)], (248 índices).
FFT-$8_{(14)}$		Tono fundamental + 4 coeficientes + 4 coeficientes delta (5 tramas). Cuantificación en 3 niveles: 1ª nivel [Tono fundamental], (2 índices). 2ª nivel [2 coeficientes], (62 índices). 3ª nivel [2 coeficientes + 4 coeficientes delta (5 tramas)], (192 índices).
FFT-$8_{(15)}$		Tono fundamental + 4 coeficientes + 4 coeficientes delta (5 tramas). Cuantificación en 3 niveles: 1ª nivel [Tono fundamental], (4 índices). 2ª nivel [2 coeficientes], (62 índices). 3ª nivel [2 coeficientes + 4 coeficientes delta (5 tramas)], (192 índices).
FFT-$8_{(16)}$		Tono fundamental + 4 coeficientes + 4 coeficientes delta (5 tramas). Cuantificación en 3 niveles: 1ª nivel [Tono fundamental], (8 índices). 2ª nivel [2 coeficientes], (56 índices). 3ª nivel [2 coeficientes + 4 coeficientes delta (5 tramas)], (192 índices).

Tabla 5.4. Conjunto de experimentos basados en el parametrizador FFT-8, empleando la información del tono fundamental y realizando el proceso de cuantificación vectorial por niveles.

5.5.2. Experimentos basados en FFT-Cepstrum con un número de coeficientes variable

a).- Procedimiento de cálculo:

1. Algoritmo transformada discreta de Fourier (radix 2 diezmado en el tiempo) con 256 muestras y desplazando 128.
2. Cálculo de 20 bandas en frecuencia siguiendo escalas de Mel.
3. Cálculo de ***n*** coeficientes siguiendo la expresión del cálculo de coeficientes MFCC.
4. Cálculo de ***n*** coeficientes Δ-cepstrum a partir de 5 tramas (*t-4, t-3, t-2, t-1 y t*).
5. El valor de ***n*** variará según el experimento: FFT-8 (10 coeficientes), FFT-$8_{(1)}$ (6 coeficientes), FFT-$8_{(2)}$ (4 coeficientes) y FFT-$8_{(3)}$ (2 coeficientes).

b).- Resultados de reconocimiento:

- El primer aspecto destacable de los resultados de reconocimiento con este grupo de experimentos es que no existe, por lo general, una relación directa entre calidad del proceso de reconocimiento y un mayor número de coeficientes empleado.
- Los resultados con la base de datos para el teléfono extendido muestran que, las mejores tasas se obtienen con 4 coeficientes (93,00%, 93,59%) y con 6 (92,26%, 94,04%). El parametrizador con 4 coeficientes cepstrales resulta ser algo mejor para las pruebas con realizaciones con modo interrogativo/exclamativo, aunque no existe un patrón de comportamiento definido en función del tipo de entrenamiento realizado.
- Por su parte, los resultados para los casos con 2 y 10 coeficientes son también similares entre sí (88,95%, 91,56%) y (87,93%, 86,19%), respectivamente, pero siempre peores que los obtenidos para los dos casos anteriormente comentados. No obstante, el parametrizador con 10 coeficientes produce mejores tasas,

especialmente para el ensayo con realizaciones con modo interrogativo/exclamativo.

- Con la base de datos del proyecto IVORY el comportamiento de los diferentes modelos de extracción de parámetros es algo diferente. En este caso, las tasas de reconocimiento del experimento base FFT-8 son muy buenas (62,87%, 61,97%, 26,40%), en especial las referidas a las pruebas con realizaciones ruidosas.

- El experimento, que presenta las mejores tasas para los ensayos con realizaciones silenciosas y de habla con *stress*, es el FFT-$8_{(1)}$ (64,50%, 67.30%, 23,60%). Sin embargo para el caso ruidoso ofrece peores prestaciones que el supuesto anterior.

- Para el resto de experimentos se observa una clara tendencia a la pérdida en poder de discriminación, según se disminuye el número de coeficientes empleados (59,90%, 59,43%, 20,20%) (53,13%, 42,27%, 10,53%).

c).- Índices de fiabilidad:

- Conservan en buena medida las tendencias presentes para las tasas de reconocimiento, aunque en este apartado las diferencias se acortan para la base de datos del proyecto IVORY.

- Para la base de datos de teléfono los mejores índices de fiabilidad se obtienen para los experimentos con 6 coeficientes o FFT-$8_{(1)}$ (<1,39- 10,85>, <1,7- 11,23>) y con 4 coeficientes o FFT-$8_{(2)}$ (<1,56- 10,85>, <1,37- 11,07>).

- Por detrás encontraríamos al experimento base (<2,00- 8,79>, <1,68- 9,05>) y al modelo con 2 coeficientes o FFT-$8_{(3)}$ (<1,81- 8,52>, <1,96- 8,57>).

- Para la base de datos del proyecto IVORY solamente el parametrizador que emplea 2 coeficientes está claramente por detrás. Para el resto las diferencias son mínimas a favor de uno u otro.

5.5.3. Experimentos basados en FFT-Cepstrum y cuantificación vectorial por niveles

a).- Procedimiento de cálculo:

1. Algoritmo transformada discreta de Fourier (radix 2 diezmado en el tiempo) con 256 muestras y desplazando 128.
2. Cálculo de 20 bandas en frecuencia siguiendo escalas de Mel.
3. Cálculo de 10 coeficientes siguiendo la expresión del cálculo de coeficientes MFCC. Para los experimentos FFT-$8_{(7)}$ FFT-$8_{(8)}$ y FFT-$8_{(9)}$ el número de coeficientes será de 4.
4. Cálculo de 10 coeficientes Δ-cepstrum a partir de 5 tramas (*t-4, t-3, t-2, t-1 y t*). Para los experimentos FFT-$8_{(7)}$ FFT-$8_{(8)}$ y FFT-$8_{(9)}$ el número de coeficientes será de 4.
5. El proceso de cuantificación vectorial se realizará empleando un número de niveles diferente según el experimento: 2 niveles (FFT-$8_{(4)}$ y FFT-$8_{(7)}$), 3 niveles (FFT-$8_{(5)}$ y FFT-$8_{(8)}$) y por último, 4 niveles (FFT-$8_{(6)}$ y FFT-$8_{(9)}$).

b).- Resultados de reconocimiento:

- Los resultados de reconocimiento para las pruebas realizadas con la base de datos del teléfono extendido muestran como la aplicación del proceso de cuantificación vectorial por niveles, si bien afecta al resultado final, no siempre significa una pérdida de capacidad de discriminación.
- En las pruebas con realizaciones de habla con modo enunciativo, la aplicación de este procedimiento se traduce en una mejora en las tasas de reconocimiento respecto al experimento base 88,93%(FFT-8). En concreto tenemos: 89,89% (FFT-$8_{(4)}$), 89,41% (FFT-$8_{(5)}$) y 91,07% (FFT-$8_{(6)}$).
- Por su parte, las pruebas con realizaciones de habla con modo interrogativo/exclamativo ofrecen el efecto contrario, es decir, la aplicación de

este procedimiento se traduce en una pérdida de capacidad discriminante, más notoria según se emplean más fases de cuantificación: 91,56% (FFT-8), 90,41% (FFT-$8_{(4)}$), 90,00% (FFT-$8_{(5)}$), 90,15% (FFT-$8_{(6)}$).

- La comparación de los resultados del experimento FFT-$8_{(2)}$ (92,26%, 94,04%) con los experimentos FFT-$8_{(7)}$ (90,59%, 92,52%), FFT-$8_{(8)}$ (88,93%, 89,26%) y FFT-$8_{(9)}$ (86,96%, 88,37%) muestra, cómo el añadir un número mayor de niveles de cuantificación empeora los resultados cuando se emplean 4 coeficientes. Sin embargo para el caso FFF-$8_{(7)}$, las tasas siguen siendo superiores a las obtenidas en el experimento base.

- Con la base de datos del proyecto IVORY los resultados son radicalmente distintos. Para este caso la pérdida de discriminación es importante en todas las situaciones respecto al caso inicial FFT-8 (62,87%, 61,97%, 26,40%): FFT-$8_{(4)}$ (54,20%, 48,37%, 10,17%), FFT-$8_{(5)}$ (55,57%, 48,80%, 12,77%) y FFT-$8_{(6)}$ (52,20%, 49,70%, 10,17%).

- La comparación de los resultados del experimento FFT-$8_{(2)}$ (59,90%, 59,43%, 20,20%) con los experimentos FFT-$8_{(7)}$ (56,47%, 51,53%, 11,00%), FFT-$8_{(8)}$ (54,37%, 48,67%, 11,67%) y FFT-$8_{(9)}$ (55,40%, 44,57%, 8,93%) muestra, al igual que con la otra base de datos, cómo al añadir un número mayor de niveles de cuantificación los resultados empeoran. Sin embargo en esta ocasión ninguna configuración de entrenamiento de este grupo proporciona mejores resultados que el experimento base (FFT-8).

c).- Índices de fiabilidad:

- Presentan un esquema análogo al comentado para las tasas de reconocimiento.

- Para la base de datos del teléfono extendido y respecto al experimento base FFT-8 (<2,00- 8,79>, <1,68- 9,05>), se obtienen unos resultados ligeramente mejores para el caso de los ensayos con modo enunciativo y algo peores para la situación de habla con entonación interrogativa y exclamativa: FFT-$8_{(4)}$ (<1,22- 8,38>, <1,84- 8,34>), FFT-$8_{(5)}$ (<1,32- 8,84>, <1,74- 8,78>) y FFT-$8_{(6)}$ (<1,18- 8,82>, <1,8- 9,06>).

- Para los experimentos con 4 coeficientes, las tasas son siempre algo peores si las comparamos con la situación de cuantificación vectorial en una sola etapa, aunque ligeramente mejores que para el experimento base: FFT-$8_{(2)}$ (<1,56- 10,85>, <1,37- 11,07>), FFT-$8_{(7)}$ (<1,01- 10,07>, <1,54- 9,99>), FFT-$8_{(8)}$ (<1,25- 8,7>, <1,77- 8,25>) y FFT-$8_{(9)}$ (<1,32- 8,31>, <1,75- 8,15>).

- Para la otra base de datos, los índices de fiabilidad están siempre por detrás de los valores de referencia FFT-8 (<2,91- 7,48>, <3,16- 7,57>, <2,65- 3,68>) y FFT-$8_{(2)}$ (<3,57- 8,03>, <3,34- 7,45>, <2,44- 2,82>). Los mejores casos, como antes, vienen representados por los experimentos con 2 niveles de cuantificación: FFT-$8_{(4)}$ (<2,83- 5,86>, <2,83- 5,61>, <2,32- 2,43>) y FFT-$8_{(7)}$ (<2,91- 6,66>, <2,66- 5,73>, <2,51- 1,69>).

5.5.4. Experimentos basados en el tono fundamental y coeficientes FFT-Cepstrum

a).- Procedimiento de cálculo:

1. Estimación del tono fundamental a partir de la señal de voz en el dominio del tiempo.
2. Algoritmo transformada discreta de Fourier (radix 2 diezmado en el tiempo) con 256 muestras y desplazando 128.
3. Cálculo de 20 bandas en frecuencia siguiendo escalas de Mel.
4. Cálculo de 10 coeficientes siguiendo la expresión del cálculo de coeficientes MFCC.
5. Cálculo de 10 coeficientes Δ-cepstrum a partir de 5 tramas (*t-4, t-3, t-2, t-1 y t*).
6. El proceso de cuantificación vectorial se realizará con un diferente número de niveles, que van desde 1 hasta 4 niveles de cálculo: 1 nivel (FFT-$8_{(10)}$), 2 niveles (FFT-$8_{(11)}$ y FFT-$8_{(14)}$), 3 niveles (FFT-$8_{(12)}$ y FFT-$8_{(15)}$) y 4 niveles (FFT-$8_{(13)}$ y FFT-$8_{(16)}$).

b).- Resultados de reconocimiento:

- Para la base de datos telefónica se observan dos fenómenos. El primero hace referencia al experimento FFT-$8_{(10)}$ (93,19%, 94,22%). Los resultados de añadir la información de tono fundamental producen una mejora respecto al experimento de partida FFT-8: (88,93%, 91,56%) y FFT-$8_{(2)}$ (92,26%, 94,04%).
- El segundo aspecto a destacar es cómo la realización de diversas fases de cuantificación produce una pérdida de la capacidad de reconocimiento respecto a los resultados que se obtienen con el caso FFT-$8_{(10)}$. En concreto: FFT-$8_{(11)}$: (87,60%, 89,78%), FFT-$8_{(12)}$: (86,11%, 87,70%), FFT-$8_{(13)}$: (80,07%, 82,07%), FFT-$8_{(14)}$: (88,44%, 87,78%), FFT-$8_{(15)}$: (84,63%, 81,52%), FFT-$8_{(16)}$: (80,26%,

79,86%).

- Con la base de datos del proyecto IVORY los resultados siguen en esencia la misma pauta. Si comparamos los experimentos base FFT-8 (62,87%, 61,97%, 26,40%) y FFT-$8_{(2)}$ (59,90%, 59,43%, 20,20%), comprobamos cómo la adición de la información del tono fundamental mejora las prestaciones del sistema hasta casi hacerlas iguales a las que nos encontramos en el caso de utilizar 10 coeficientes (FFT8). En concreto: FFT-$8_{(10)}$ (62,63%, 60,90%, 19,40%). Solamente el ensayo con realizaciones ruidosas experimenta una caída en las tasas de reconocimiento.

- Para el resto de experimentos, como antes, la realización de la cuantificación vectorial por niveles afecta negativamente a los resultados. A veces esta pérdida es muy importante como para el ensayo FFT-$8_{(16)}$ (42,30%, 41,13%, 9,60%).

c).- Índices de fiabilidad:

- Los índices de fiabilidad para el primer conjunto de datos (teléfono extendido) muestran cómo el experimento que incluye la información de tono fundamental (FFT-$8_{(10)}$) es algo mejor que los casos de base: FFT-$8_{(10)}$ (<1,57- 10,97>, <1,32- 11,27>), FFT-8 (<2,00- 8,79>, <1,68- 9,05>) y FFT-$8_{(2)}$ (<1,56- 10,85>, <1,37- 11,07>).

- Con la otra base de datos se repiten las conclusiones citadas en el punto de resultados de reconocimiento, es decir, sus índices (FFT-$8_{(10)}$) se encuentra casi a la altura de los del caso base inicial FFT-8. Concretamente encontramos los siguientes resultados: FFT-8 (<2,91- 7,48>, <3,16- 7,57>, <2,65- 3,68>), FFT-$8_{(2)}$ (<3,57- 8,03>, <3,34- 7,45>, <2,44- 2,82>) y FFT-$8_{(10)}$ (<3,25- 7,49>, <3,48- 7,67>, <3,53- 2,79>).

- El resto de experimentos FFT-$8_{(11)}$ al FFT-$8_{(16)}$ presentan peores resultados, que cualquiera los tres casos comentados hasta ahora, para la inmensa mayoría de situaciones. En cualquier caso, ninguno de estos índices mejora al ganador de la terna anterior con independencia de la categoría analizada.

5.5.5. Análisis de la medida de separación entre clases

En esta sección comentaremos algunas ideas respecto a las medidas de separación de clases, que encontramos al realizar este grupo de nuevos experimentos.

La motivación principal para realizar este análisis es comprobar la viabilidad de este tipo de medida, cuando se aplica a sistemas de reconocimiento de voz como forma de predecir su capacidad de discriminación.

Si analizamos el experimento base (FFT-8), se comprueba cómo para las dos bases de datos los coeficientes que aportan mayor separación son los primeros estáticos (*c2*, *c1*, *c4*) y el primero de los dinámicos o *dc1* (Figura 5.6 y Figura 5.7). El coeficiente c6 para la base de datos del teléfono extendido estaría también incluido en el primer grupo.

La mayor importancia de los primeros coeficientes se ve aumentada cuando se reduce el número de parámetros de los experimentos: 6 (FFT-$8_{(1)}$), 4 (FFT-$8_{(2)}$) y 2 (FFT-$8_{(3)}$). Los dos primeros experimentos mantienen, en lo esencial, el orden de los componentes mientras que en el último caso (Figura 5.12 y Figura 5.13), los coeficientes dinámicos *dc1* y *dc2* pasan a ocupar los primeros puestos. Este cambio en las posiciones podría indicar que se han eliminado algunos rasgos/coeficientes importantes.

Los experimentos FFT-$8_{(4)}$ y FFT-$8_{(5)}$ reflejan un cambio radical. El establecimiento de niveles de cuantificación vectorial se refleja en una pérdida de la capacidad de separación para estas componentes, como consecuencia de que la participación de los primeros coeficientes cepstrales se restringe a los primeros niveles.

Este fenómeno es normal, aunque el objetivo debiera ser igualar la capacidad discriminatoria entre todos los elementos. Esto indicaría para todos estos experimentos, que la separación en niveles ha sido muy drástica. Una explicación sencilla al fenómeno sería que no es conveniente separar aquellas componentes que se encuentren relacionadas entre sí.

El experimento resultante de añadir al vector de parámetros del experimento base, la información del tono fundamental (FFT-$8_{(10)}$) presenta un claro predominio de esta característica frente al resto de coeficientes (Figura 5.18 y Figura 5.19). Esto vendría a ser un reflejo de la importancia que tiene y que debería traducirse en una mejora en el funcionamiento del sistema de reconocimiento asociado.

Por el contrario, esta característica se pierde cuando se aplican, a este último caso, varios niveles de cuantificación y además se emplean pocos índices al rasgo del tono fundamental (experimentos FFT-$8_{(11)}$ al FFT-$8_{(16)}$). Aquí la conclusión sería que emplear un número de clases pequeño para la componente asociada al tono fundamental, esto es, usar el tono fundamental como indicador de trama sonora/sorda no es adecuado si se presenta junto a los cuatro primeros coeficientes cepstrales (estáticos y dinámicos) en un mismo vector de características.

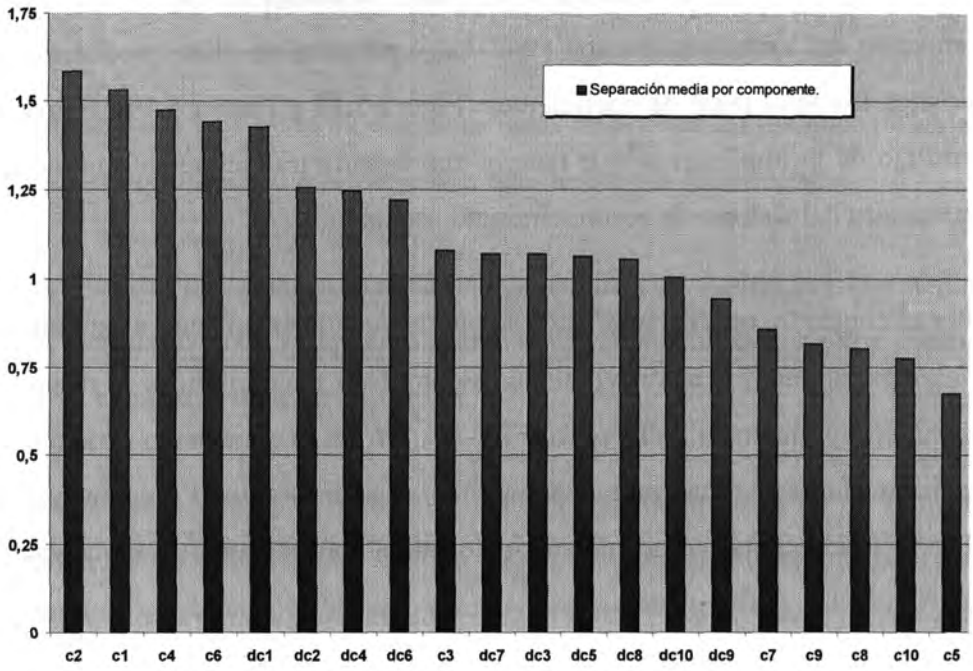

Figura 5.6. Separación media de clases para cada componente del vector de parámetros del experimento FFT-8a (FFT-Cepstrum [10 coeficientes + 10 coeficientes delta <5 tramas> + 10 coeficientes delta-delta + energía trama + energía trama diferencia <2 tramas>]).

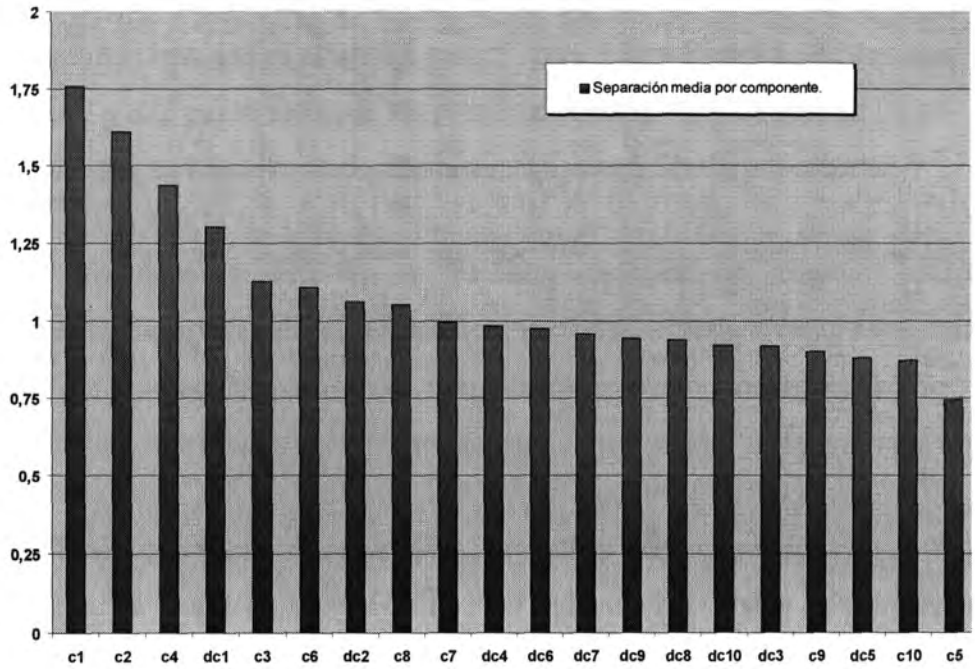

Figura 5.7. Separación media de clases para cada componente del vector de parámetros del experimento FFT-8b (FFT-Cepstrum [10 coeficientes + 10 coeficientes delta <5 tramas> + 10 coeficientes delta-delta + energía trama + energía trama diferencia <2 tramas>]).

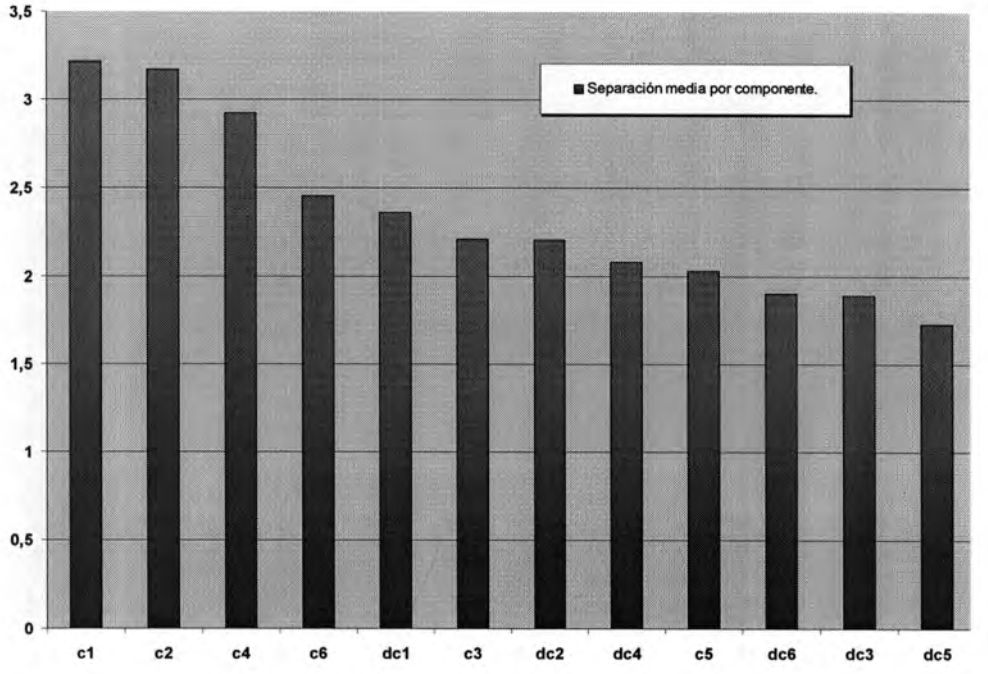

Figura 5.8. Separación media de clases para cada componente del vector de parámetros del experimento FFT-$8_{(1)}$a (FFT-Cepstrum [6 coeficientes + 6 coeficientes delta <5 tramas>]).

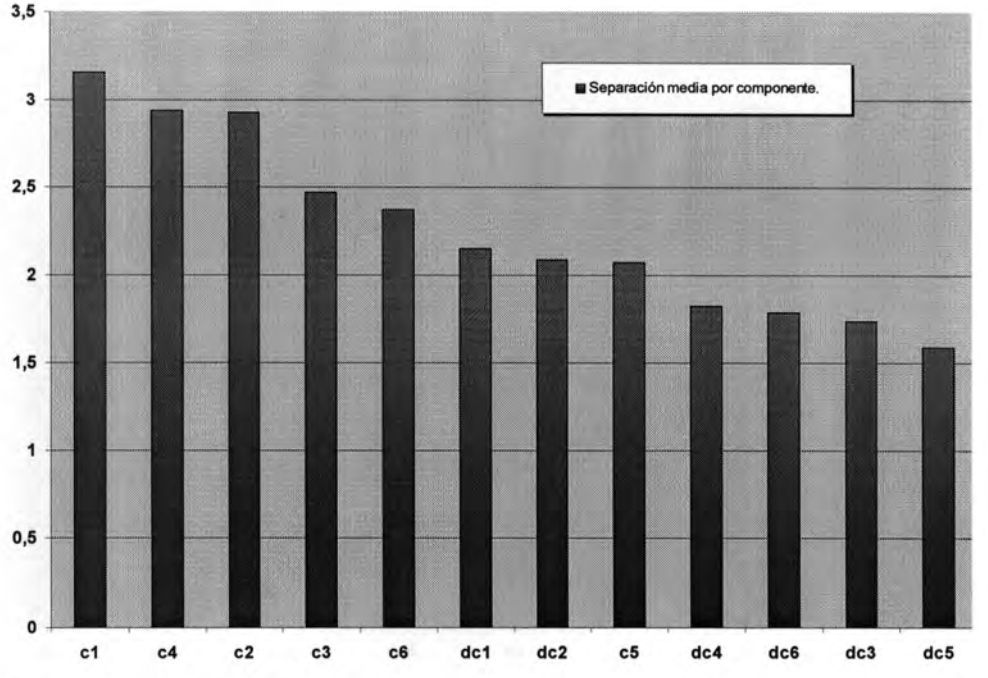

Figura 5.9. Separación media de clases para cada componente del vector de parámetros del experimento FFT-$8_{(1)}$b (FFT-Cepstrum [6 coeficientes + 6 coeficientes delta <5 tramas>]).

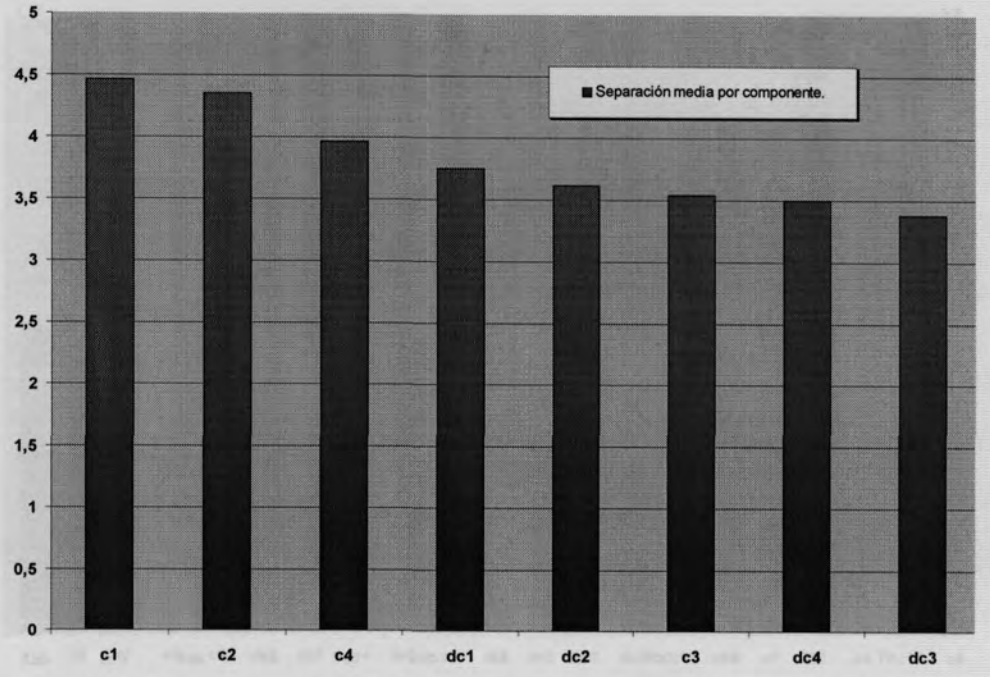

Figura 5.10. Separación media de clases para cada componente del vector de parámetros del experimento FFT-$8_{(2)}$a (FFT-Cepstrum [4 coeficientes + 4 coeficientes delta <5 tramas>]).

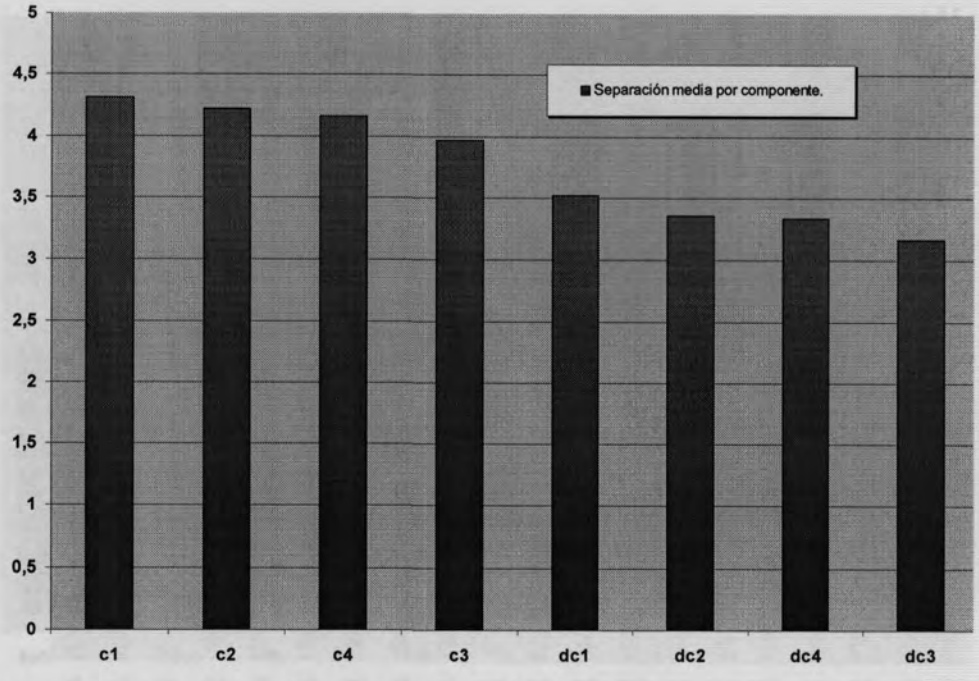

Figura 5.11. Separación media de clases para cada componente del vector de parámetros del experimento FFT-$8_{(2)}$b (FFT-Cepstrum [4 coeficientes + 4 coeficientes delta <5 tramas>]).

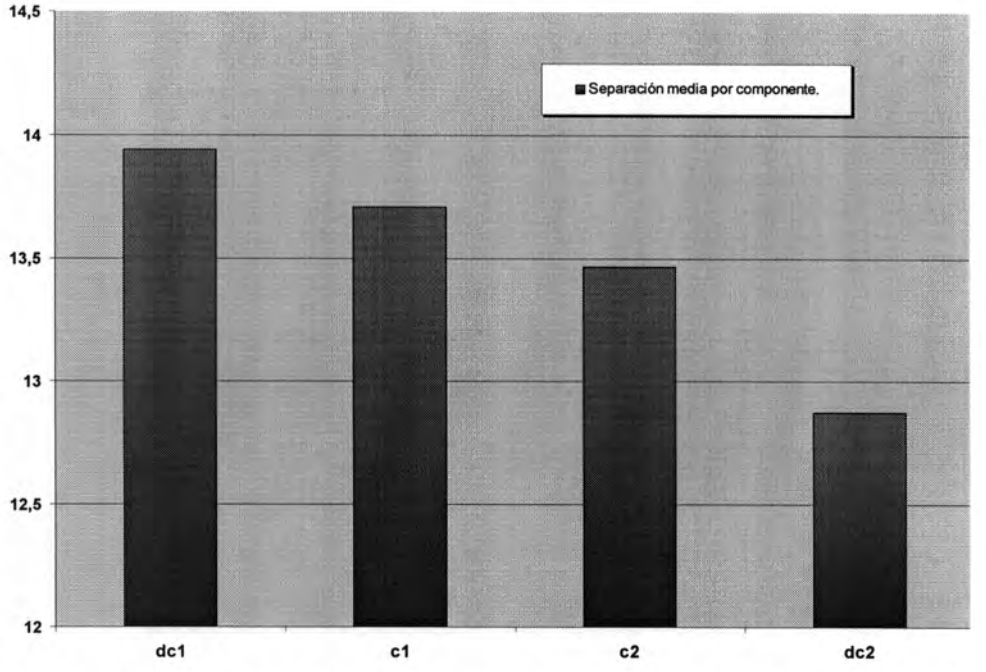

Figura 5.12. Separación media de clases para cada componente del vector de parámetros del experimento FFT-$8_{(3)}$a (FFT-Cepstrum [2 coeficientes + 2 coeficientes delta <5 tramas>]).

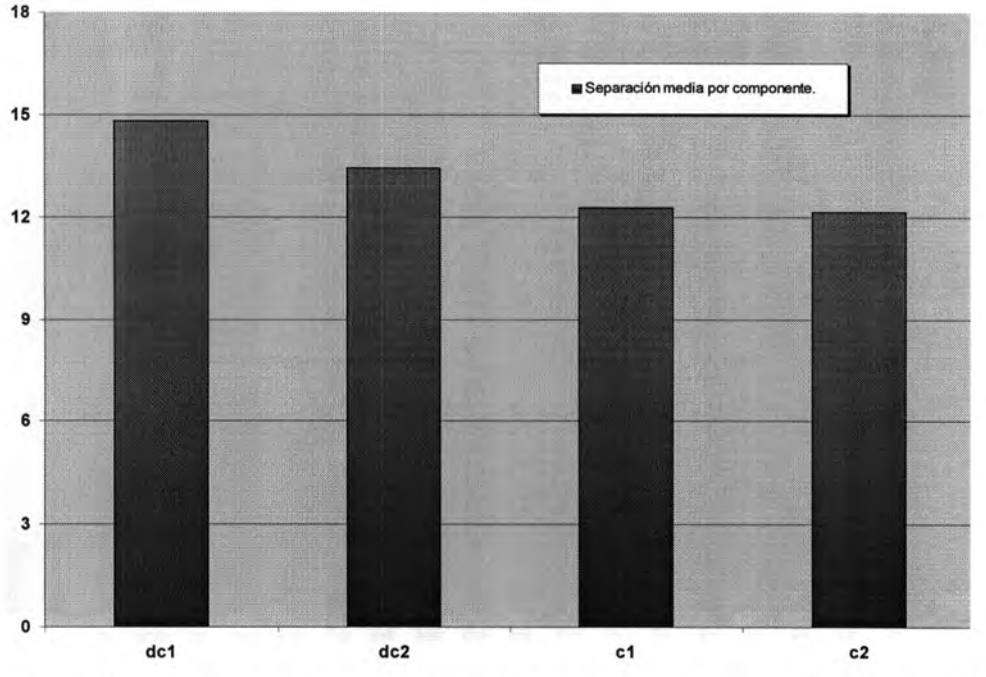

Figura 5.13. Separación media de clases para cada componente del vector de parámetros del experimento FFT-$8_{(3)}$b (FFT-Cepstrum [2 coeficientes + 2 coeficientes delta <5 tramas>]).

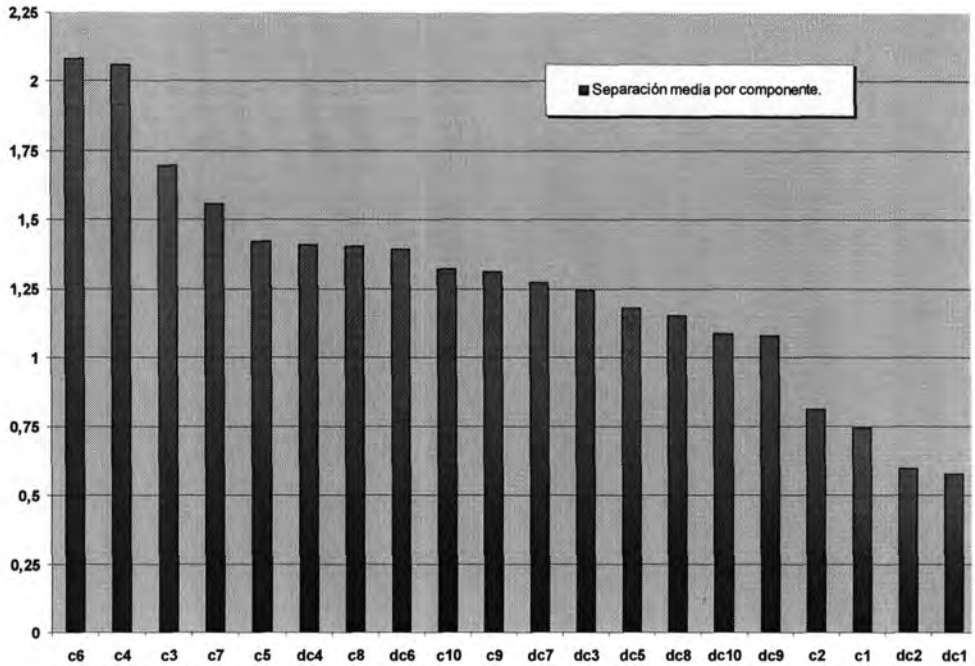

Figura 5.14. Separación media de clases para cada componente del vector de parámetros del FFT-$8_{(4)}$a (FFT-Cepstrum [10 coeficientes + 10 coeficientes delta <5 tramas>]. Cuantificación en 2 niveles: nivel 1° <16 índices> [6 coeficientes + 6 coeficientes delta], nivel 2° <240 índices> [4 coeficientes + 4 coeficientes delta]).

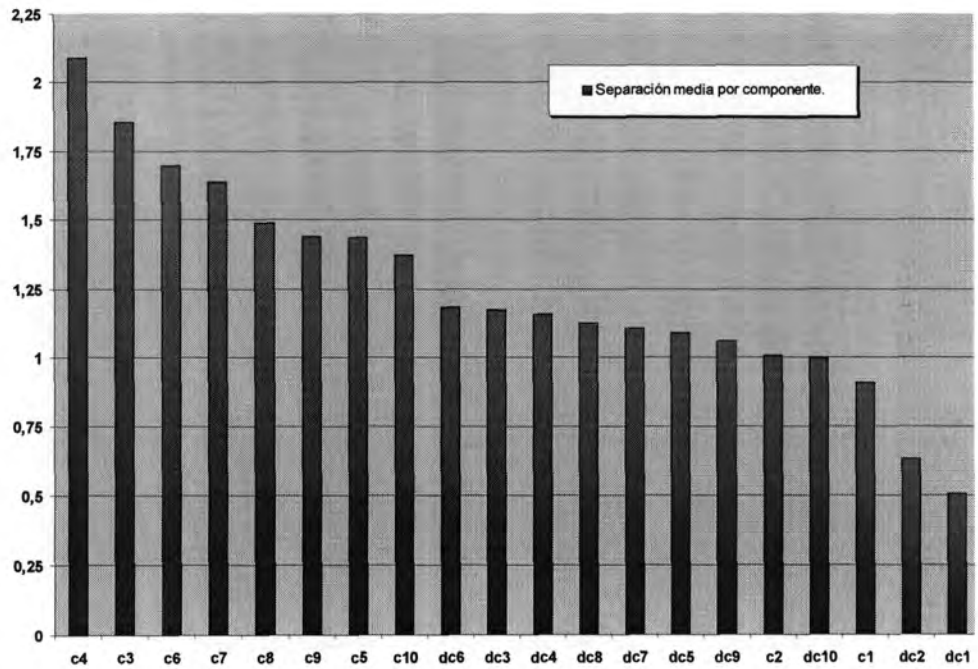

Figura 5.15. Separación media de clases para cada componente del vector de parámetros del experimento FFT-$8_{(4)}$b (FFT-Cepstrum [10 coeficientes + 10 coeficientes delta <5 tramas>]. Cuantificación en 2 niveles: nivel 1° <16 índices> [6 coeficientes + 6 coeficientes delta], nivel 2° <240 índices> [4 coeficientes + 4 coeficientes delta]).

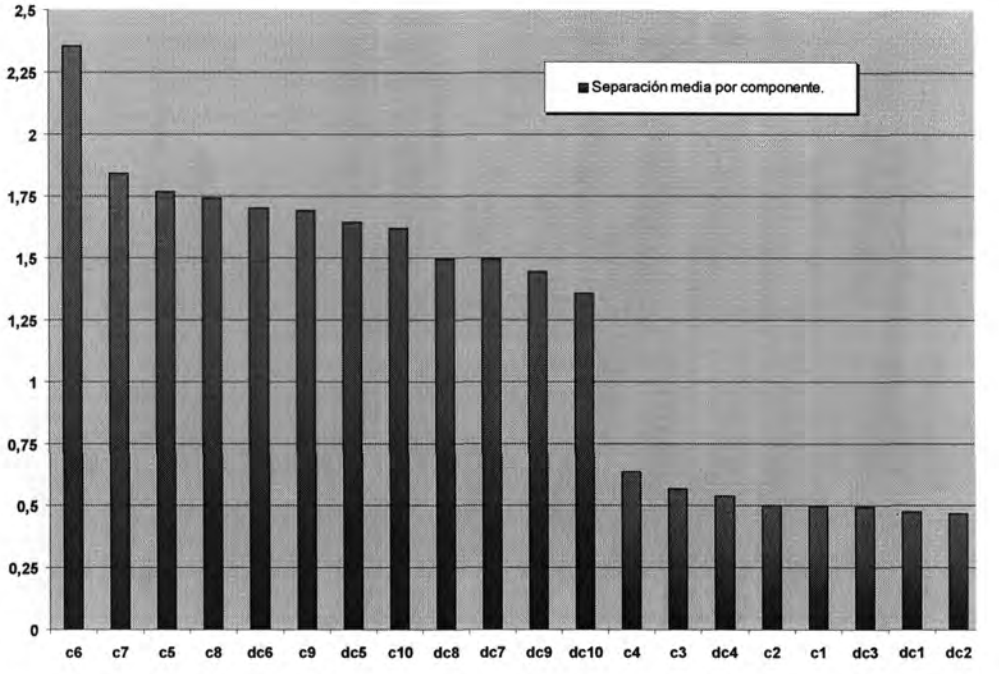

Figura 5.16. Separación media de clases para cada componente del vector de parámetros del experimento FFT-$8_{(5)}$a (FFT-Cepstrum [10 coeficientes + 10 coeficientes delta <5 tramas>]. Cuantificación en 3 niveles: nivel 1° <16 índices> [4 coeficientes + 4 coeficientes delta], nivel 2° <48 índices> [2 coeficientes + 2 coeficientes delta], nivel 3° <192 índices> [4 coeficientes + 4 coeficientes delta]).

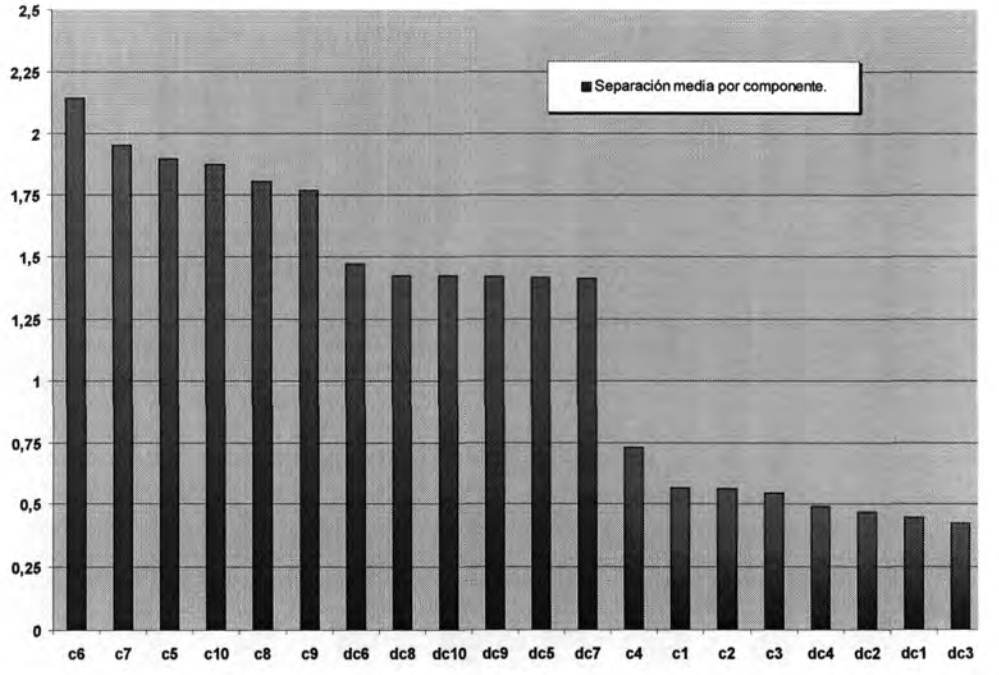

Figura 5.17. Separación media de clases para cada componente del vector de parámetros del experimento FFT-$8_{(5)}$b (FFT-Cepstrum [10 coeficientes + 10 coeficientes delta <5 tramas>]. Cuantificación en 3 niveles: nivel 1° <16 índices> [4 coeficientes + 4 coeficientes delta], nivel 2° <48 índices> [2 coeficientes + 2 coeficientes delta], nivel 3° <192 índices> [4 coeficientes + 4 coeficientes delta]).

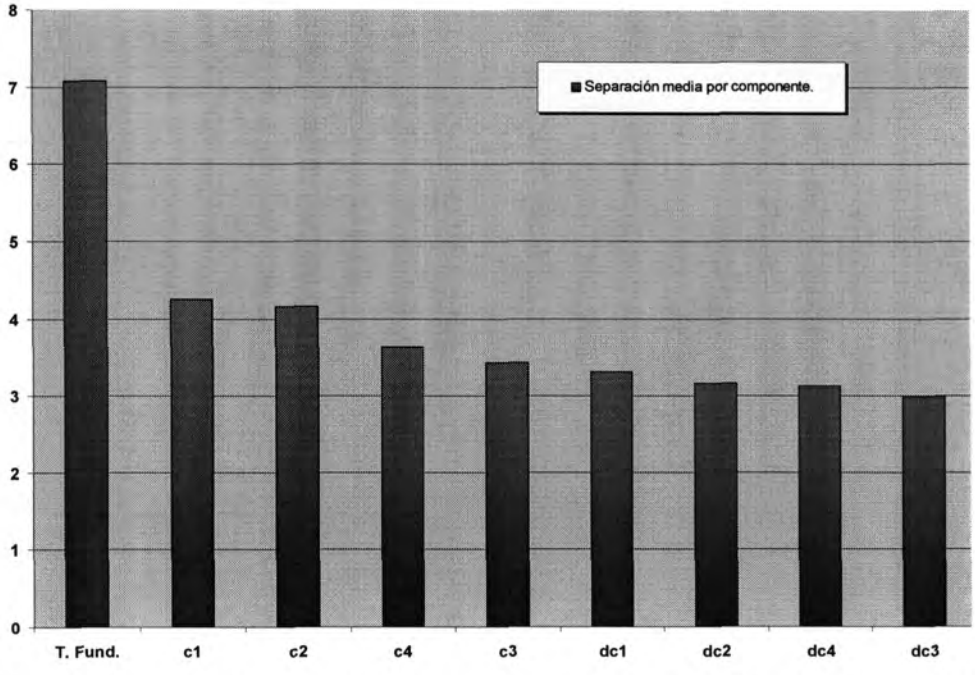

Figura 5.18. Separación media de clases para cada componente del vector de parámetros del experimento FFT-$8_{(10)}$a (Tono fundamental + FFT-Cepstrum [4 coeficientes + 4 coeficientes delta <5 tramas>]).

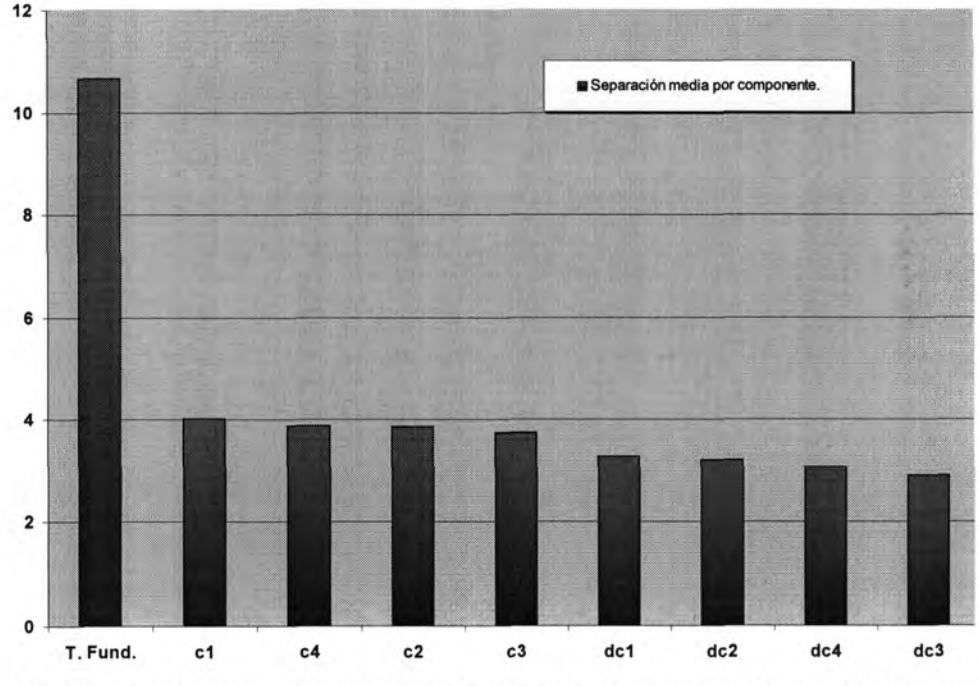

Figura 5.19. Separación media de clases para cada componente del vector de parámetros del experimento FFT-$8_{(10)}$b (Tono fundamental + FFT-Cepstrum [4 coeficientes + 4 coeficientes delta <5 tramas>]).

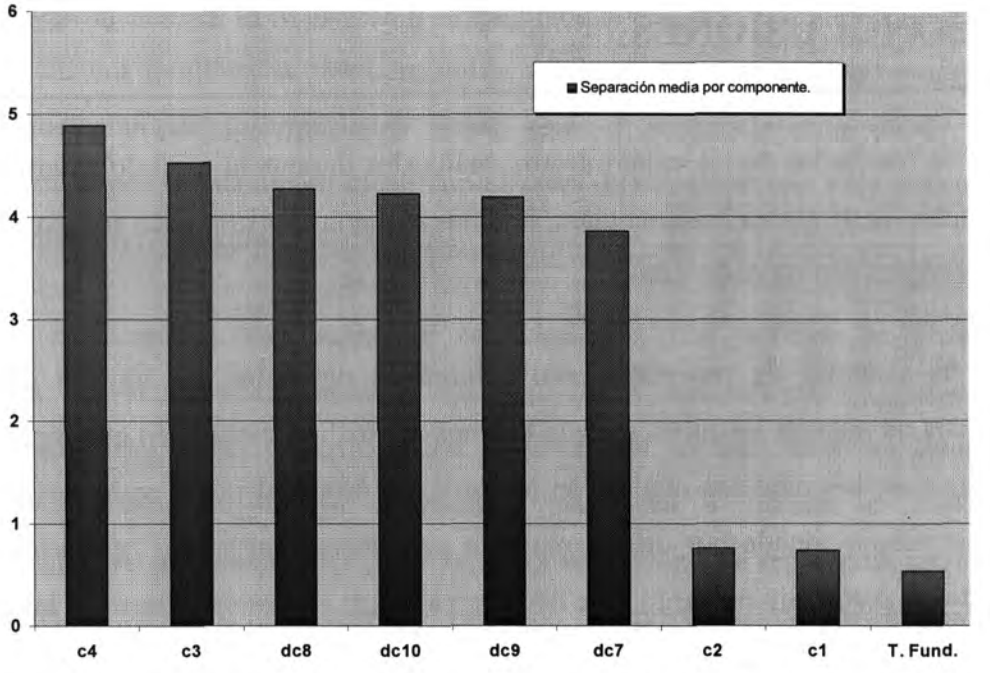

Figura 5.20. Separación media de clases para cada componente del vector de parámetros del experimento FFT-$8_{(15)}$a (Tono fundamental + FFT-Cepstrum [4 coeficientes + 4 coeficientes delta <5 tramas>]. Cuantificación en 3 niveles: nivel 1º <4 índices> [Tono fundamental], nivel 2º <60 índices> [2 coeficiente], nivel 3º <192 índices> [2 coeficientes + 4 coeficientes delta]).

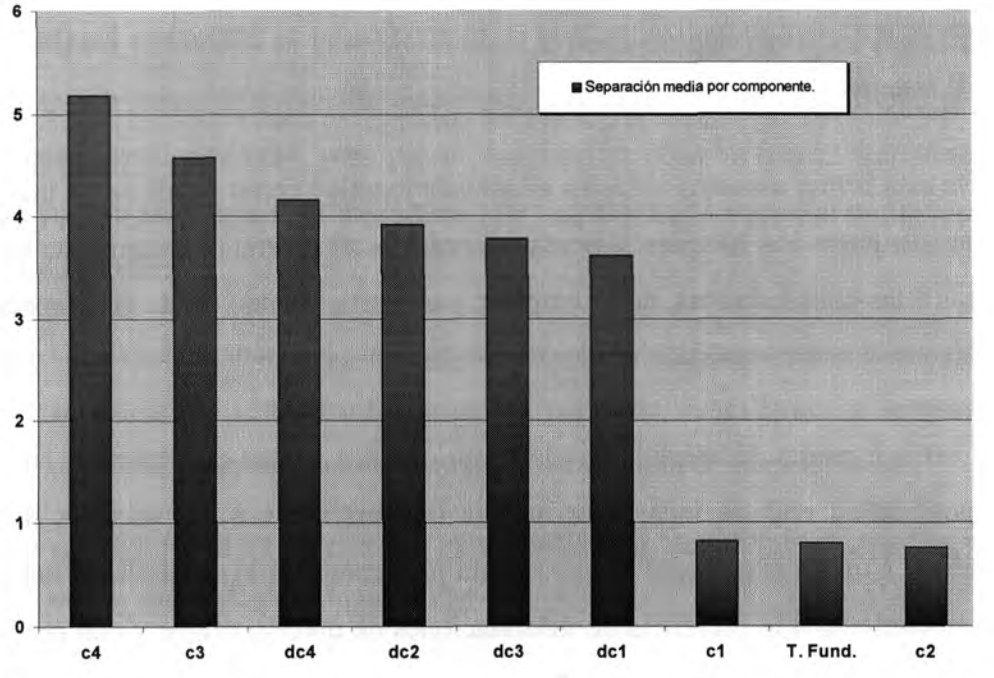

Figura 5.21. Separación media de clases para cada componente del vector de parámetros del experimento FFT-$8_{(15)}$b (Tono fundamental + FFT-Cepstrum [4 coeficientes + 4 coeficientes delta <5 tramas>]. Cuantificación en 3 niveles: nivel 1º <4 índices> [Tono fundamental], nivel 2º <60 índices> [2 coeficiente], nivel 3º <192 índices> [2 coeficientes + 4 coeficientes delta]).

5.6. CONCLUSIONES

Los resultados de los experimentos analizados durante el capítulo anterior a éste, dejan entrever de forma bastante clara, la pobreza del análisis fonético realizado por los reconocedores de voz de hoy en día.

Los sistemas de reconocimiento automático del habla, de hoy en día, están enfocados de manera inequívoca hacia la etapa de análisis estadístico que representa la tecnología de los modelos ocultos de Markov. La principal razón es la flexibilidad y potencia proporcionada por dichos modelos para incorporar nuevo "conocimiento" a partir de los datos. Sin embargo, este hecho condiciona en exceso al resto de las etapas.

El análisis fonético presente en los sistemas de reconocimiento se reduce, más o menos, a un tratamiento secuencial de la señal de voz por pequeños fragmentos de tamaño fijo. La principal virtud de este enfoque es la sencillez, puesto que el tratamiento es uniforme en todos los casos sin ningún tipo de distinción. Incluso en partes donde no hay presencia de voz se realiza, aunque en la práctica no se propaguen los resultados al resto del sistema.

De esta forma aspectos clave en el análisis fonético como puede ser el tratamiento diferenciado entre los bloques sonoros y sordos o el diferente manejo de las zonas vocálicas y las consonánticas, no se emplean para nada, cuando desde la perspectiva del conocimiento fonético-acústico se muestran como básicos.

Con este marco de trabajo, uno de los primeros aspectos que debemos considerar, es la posibilidad real de incorporar nuevas informaciones a los vectores de rasgos disponibles. Con otras palabras lo que se está planteando es la sensibilidad del proceso de reconocimiento a la presencia de informaciones de diverso origen y con propiedades diferentes. Las conclusiones respecto a este punto que hemos tratado en el presente capítulo son:

La inclusión de nuevas componentes al vector de rasgos es potencialmente dañina. Esto se verifica en que incluso para el caso de vectores con informaciones uniformes (ej.

coeficientes cepstrales), la reducción o la ampliación, según se quiera ver, de parte de las componentes modifica sustancialmente la respuesta del sistema. Esto quiere decir, que la inclusión de informaciones útiles para el reconocimiento puede quedar contrarrestada por la incompatibilidad con el resto de componentes. Este podría ser el caso de incluir información de formantes.

La problemática de incorporar conocimientos con niveles de abstracción diferentes en vectores de tamaño fijo (ej. rasgo sonoro/sordo y rasgos para determinación del punto de articulación en oclusivas sordas), tanto en número de componentes como en tiempo de análisis, nos llevó a plantear el método de cuantificación vectorial jerárquica controlado por una medida de separación entre clases. Los resultados vienen a demostrar que este tipo de aproximación, por lo general, no mejora los resultados sino mas bien al contrario, pero que a pesar de todo puede ser un planteamiento útil para afrontar este problema.

Por último, se planteó un caso real de incorporar una medida del tono fundamental al vector de características. La utilización del rasgo del tono fundamental en sistemas de reconocimiento [KOM97] es bastante inusual, a pesar de que desde un punto de vista fonético se vea como muy útil. En este caso, la mejora es apreciable frente al mismo experimento sin su concurso, pero por el contrario se observa que el comportamiento empeora cuando añadimos el procedimiento de cuantificación vectorial por niveles. Esto significa dos cosas:

a).- La idea de emplear el valor del tono fundamental como un detector del rasgo sonoro/sordo (pocos índices en un nivel de cuantificación separado), no parece ser muy compatible con el resto del vector de características.

b).- La separación en niveles de cuantificación de los diferentes componentes del vector resulta ser también una labor delicada.

Como resumen de este apartado podemos decir, que si bien, la incorporación de nuevos conocimientos útiles es potencialmente beneficiosa para los sistemas de reconocimiento del habla, la propia configuración de dichos sistemas puede dificultar seriamente dichos intentos.

6. CONCLUSIONES

6.1. PRINCIPALES CONCLUSIONES

La principal conclusión que se obtiene del trabajo realizado es la confirmación de nuestra hipótesis inicial, es decir, que la aportación de conocimiento directo, proveniente del campo de la fonética acústica, al problema del reconocimiento automático de la voz, en concreto a la etapa de extracción de características, puede constituir una base sólida con la que poder analizar el comportamiento y capacidad de discriminación de dichos sistemas, así como una forma de mejorar sus prestaciones.

La inclusión de conocimiento en los sistemas de reconocimiento es potencialmente provechosa, sin embargo, la arquitectura y configuración de los sistemas actuales de reconocimiento del habla dificulta enormemente dicha aplicación.

Por otra parte, los estudios de fonética acústica carecen de la profundidad y generalidad suficientes como para que dicho proceso de incorporación sea coherente y completo. Esto es especialmente notorio para el caso del castellano.

El análisis fonético o de extracción de rasgos de un reconocedor de voz debería permitir realizar el examen de la señal de voz sin limitaciones en cuanto al tamaño de los fragmentos de análisis. No obstante y por aspectos de compatibilidad con el resto de las etapas de los referidos sistemas, las tramas de información podrían generarse a intervalos regulares de tiempo.

Así mismo un examen fonético-acústico adecuado debería distinguir claramente el tratamiento de zonas vocálicas y consonánticas, apoyando el análisis de estas últimas en las primeras. De igual forma, el estudio de regiones sonoras y sordas ha de ser claramente diferenciado.

A pesar de la limitación anterior y con independencia de la técnica de tratamiento de la señal empleada, los mecanismos de extracción de características mejoran notablemente si se incorporan parámetros dinámicos al vector de componentes espectrales.

La combinación de las informaciones presentes en las diferentes bandas o regiones del espectrograma y la aplicación de escalas no lineales en todo el rango de frecuencias, garantizan también los mejores resultados de reconocimiento.

La comparación de los resultados para los diferentes grupos de técnicas de análisis permite concluir:

Ningún esquema de obtención de rasgos, por sí solo, resulta ser el mejor en todas las situaciones.

La construcción o entrenamiento de los sistemas con conjuntos de datos adaptados a las condiciones de funcionamiento real, así como, a los usuarios finales, es un factor clave para garantizar la calidad del reconocedor de voz.

En este sentido la incorporación equilibrada de realizaciones de personas de diferente género, con variantes dialectales distintas y edades diversas, son factores a tener en cuenta cuando se desarrollan estos sistemas.

El volumen de datos disponibles para realizar el entrenamiento es en la práctica el parámetro que más influye en la calidad del resultado final. Fundamentalmente encontramos dos razones:

1. En los reconocedores actuales, la forma usual de enfrentarse a las múltiples fuentes de variabilidad presentes en el proceso de la comunicación pasa por incorporar realizaciones, que incluyan las "distorsiones" introducidas por dichas fuentes de variabilidad, al conjunto de datos empleados durante la construcción del sistema.

2. Los modelos de reconocimiento, por regla general, no incluyen ningún tipo de análisis fonológico, que permita asociar realizaciones de habla diversas en cuanto a su contenido fonético. En la práctica, este conocimiento aparece únicamente recogido de forma implícita en los datos de entrenamiento.

3. Muchos de los métodos prevalecientes dentro del campo del reconocimiento automático de discurso, son aquellos que permiten la mejora de los sistemas

cuando la disponibilidad de datos para el entrenamiento aumenta.

Los sistemas de reconocimiento de hoy en día se fundamentan en las operaciones de autómatas probabilísticos, que no permiten extraer medidas fiables acerca de la calidad de sus respuestas. En este sentido y con independencia de las tasas de reconocimiento que se puedan llegar a obtener, la confianza que acompaña a las decisiones del reconocedor es prácticamente nula.

El módulo de examen fonético, en conjunción con las etapas de análisis fonológico y sintáctico, debe proveer mecanismos eficaces para la elaboración y evaluación de segundas hipótesis, que además podrían servir como forma de adaptación a las características de habla de diferentes locutores.

6.2. APORTACIONES ORIGINALES

En este apartado se resumen las aportaciones de la tesis que resultan novedosas frente a los trabajos realizados en el área hasta el momento. Las aportaciones originales se basan en los resultados más interesantes conseguidos mediante los diversos estudios experimentales llevados a cabo.

En primer lugar, la realización de un exhaustivo análisis de las capacidades de discriminación de los diferentes modelos de obtención de rasgos fonéticos presentes en los actuales sistemas de reconocimiento automático del habla en condiciones que garantizan la homogeneidad de datos utilizada en los análisis de contraste.

La creación de un algoritmo de cálculo del tono fundamental basado en técnicas de programación dinámica, que se ha empleado como rasgo adicional en diversos modelos de obtención de características.

El estudio de la posible aplicabilidad de nuevas fuentes de información acústica a los vectores de parámetros clásicos y de su compatibilidad con el resto de etapas de los sistemas de reconocimiento de discurso.

La implementación en una plataforma DSP, en concreto TMS320C31-60, de dos diferentes versiones del reconocedor de voz de palabras aisladas.

Aunque la incorporación de conocimiento fonético-acústico a los citados sistemas no es completa, ni mucho menos, sí que se puede considerar como referencia innovadora para el desarrollo de trabajos más elaborados que se basen en las ideas, métodos y herramientas originales empleados en la tesis.

La aplicación informática, producida como punto de partida de los trabajos de investigación, puede ser empleada por cualquier estudioso interesado en el campo del reconocimiento de voz. Sus principales ventajas son su funcionamiento sin necesidad de un hardware específico y la estructura modular y abierta, donde es posible modificar o ampliar cualquier funcionalidad que se considere adecuada, así como, reemplazar de forma sencilla aquellas etapas de los sistemas de reconocimiento que se deseen.

6.3. LÍNEAS FUTURAS DE CONTINUACIÓN

Las posibilidades de ampliación y continuación de los trabajos desarrollados a lo largo de la presente tesis se dividen en dos áreas fundamentales:

1. Campo del reconocimiento automático de la voz, donde resultaría muy útil la incorporación de una etapa de análisis fonético y fonológico elaborada.
2. Disciplinas relacionadas con los estudios fonéticos, donde se deberían realizar estudios más profundos y ambiciosos, con el fin de poder aplicar de una manera coherente, conocimiento de estas áreas a los sistemas de reconocimiento de la voz.

En especial resultaría muy adecuado extender las investigaciones y desarrollos en los siguientes aspectos:

- Creación de una base de datos completa y equilibrada, que pudiera servir de base para posteriores estudios tanto del campo del reconocimiento del habla, como de fonética acústica.
- Caracterización acústica de vocales en habla continua, recogiendo todas aquellas fuentes de variabilidad que producen desviaciones en su modelo base: coarticulación, ritmo de habla, género del hablante, edad, etc. Este ambicioso análisis debería permitir la realización de un sistema reconocedor, que estuviese fundamentado en la detección fiable de zonas vocálicas.
- Caracterización acústica de consonantes, diferenciando claramente los elementos deducibles por medio de transiciones vocálicas adyacentes y aquellos otros deducibles de forma independiente.
- Realización de sistemas de reconocimiento, que incorporen además de la etapa de examen fonético, un módulo con un análisis fonológico.

7. BIBLIOGRAFÍA

[ABD98] A. M. Abdelatty, J. Van der Spiegel and P. Mueller, "An Acoustic-Phonetic Feature-Based System for the Automatic Recognition of Fricative Consonants", *Proc. of ICASSP'98*, Seattle, Estados Unidos, 12-15 mayo 1998, pp. 961-964.

[ABE88] M. Abe et al., "Voice Conversion Through Vector Quantization", Proc. of ICASSP'88, Nueva York, Estados Unidos, 11-14 abril 1988, pp. 655-658.

[ABE91] M. Abe, "A Segmental-Based Approach to Voice Conversion", Proc. of ICASSP'91, Toronto, Canadá, 14-17 mayo 1991, pp. 765-768.

[ACE90] A. Acero and R. M. Stern, "Environmental Robustness in Automatic Speech Recognition", *Proc. of ICASSP'90*, Alburquerque, Estados Unidos, 3-6 abril 1990, pp. 849-852.

[AGA97] H. Agaiby et al., "Commercial Speech Recognizers Performance under Adverse Conditions, A Survey", Proc. of the ESCA-NATO Tutorial and Research Workshop on Robust Speech Recognition for Unknown Communication Channels, Pont-à-Mousson, Francia, 17-18 abril 1997, pp. 163-166.

[AGU95] L. Aguilar, M. Machuca, "Pragmatic Factors Affecting the Phonetic Properties of Diphthongs", *Proc. of EUROSPEECH'95, Madrid, septiembre* 1995, pp. 2251-2254.

[ALA91] E. Alarcos, *Fonología española*, Editorial Gredos, Madrid, 1991.

[ALE76] A. D. Aleksandrov et al. *La Matemática: su Contenido, Métodos y Significado*, Alianza Universidad, 1976

[ALL87] M. Allerhand, *Knowledge-Based Speech Pattern Recognition*, Kogan Page LTD., 1987.

[ALL93] F. Alleva, X. Huang and M. Y. Hwang, "An Improved Search Algorithm Using Incremental Knowledge for Continuous Speech Recognition", *Proc. of ICASSP'93*, Minneapolis, Estados Unidos, 27-30 abril 1993, Vol. II, pp. 307-310.

[ALV90] V. M. Alvarado and H. F. Silverman, "Experimental Results Showing the Effects of Optimal Spacing Between Elements of a Linear Microphone Array", *Proc. of ICASSP'90*, Alburquerque, Estados Unidos, 3-6 abril 1990, pp. 837-840.

[ANA94] A. Anastasakos et al., "Adaptation to New Microphones Using Tied-Mixture Normalization", *Proc. of ICASSP'94*, Adelaida, Australia, 19-22 abril 1994, Vol. I, pp. 433-436.

[ANA95] A. Anastasakos, R. Schwartz and H. Shu, "Duration Modeling in Large Vocabulary Speech Recognition", *Proc. of ICASSP'95*, Detroit, Estados Unidos, 9-12 mayo 1995, Detroit, Estados Unidos, 9-12 mayo 1995, pp. 628-631.

[APP89] T. H. Applebaum and B. A. Hanson, "Enhancing the Discrimitation of Speaker Independent Hidden Markov Models with Corrective Training", *Proc. of ICASSP'89*, Glasgow, Reino Unido, 23-26 mayo 1989, pp. 302-305.

[ASS96] P. F. Assmann, "Modeling the perception of concurrent vowels: Role of formant transitions", *Journal of Acoustic Society of America*, Vol. 100, N° 2, agosto 1997, pp. 1141-1152.

[ATA83] B. S. Atal, "Efficient Coding of LPC Parameters by Temporal Decomposition", *Proc. of ICASSP'83*, Boston, Estados Unidos, 14- 16 abril 1983, pp. 81-84.

[AUB94] X. Aubert et al., "Large Vocabulary Continuous Speech Recognition of Wall Street Journal Data", *Proc. of ICASSP'94*, Adelaida, Australia, 19-22 abril 1994, Vol. II, pp.129-132.

[AVE87] A. Averbuch et al., "Experiments with the Tangora 20,000 Word Speech Recognizer", *Proc. of ICASSP'87*, Dallas, Estados Unidos, 6-9 abril 1987, pp. 701-704.

[BAB95] C. Baber and J. Noyes, "Automatic Speech Recognition Systems: effect of environmental stressors", *ESCA-NATO Tutorial and Research Workshop on Speech under Stress*, Lisboa, 14-15 septiembre 1995, pp. 37-40.

[BAH75] L. R. Bahl, "Decoding for Channels with Insertions, Deletions, and Substitutions with Applications to Speech Recognition", *IEEE Transactions on Information Theory*, Vol. IT-21, N°. 4, julio 1975, pp. 404- 411.

[BAH83] L. R. Bahl, F. Jelinek and R. L. Mercer, "A Maximum Likehood Approach to Continuous Speech Recognition", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. PAMI-5, N°. 2, marzo 1983, pp. 179-190.

[BAH86] L. R. Bahl et al., "Maximum Mutual Information Estimation of Hidden Markov Model Parameters for Speech Recognition", *Proc. of ICASSP'86*, pp. 49-52.

[BAH88a] L. R. Bahl et al., "Acoustic Markov Models Used in the Tangora Speech Recognition System", *Proc. of ICASSP'88*, Nueva York, Estados Unidos, 11-14 abril 1988, pp. 497-500.

[BAH88b] L. R. Bahl et al., "A New Algorithm for the Estimation of Hidden Markov Model Parameters", *Proc. of ICASSP'88*, Nueva York, Estados Unidos, 11-14 abril 1988, pp. 493-496.

[BAH91] L. R. Bahl et al., "Decision Trees for Phonological Rules in Continuous Speech", *Proc. of ICASSP'91*, Toronto, Canadá, 14-17 mayo 1991, pp. 185-188.

[BAH94] L. R. Bahl et al., "Robust Methods for Using Context-Dependent Features and Models in a Continuous Speech Recognizer", *Proc. of ICASSP'94*, Adelaida, Australia, 19-22 abril 1994, Vol. I, pp. 533-536.

[BAK75] J. K. Baker, "The DRAGON System- An Overview", *IEEE Transactions on Acoustic, Speech and Signal Processing*, Vol. ASSP-23, N°. 1, febrero 1975, pp. 24-29.

[BAU70] L. E. Baum et al., "A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains", *The Annals of Mathemartical Statistics*, Vol. 41, N°. 1, 1970, pp. 164-171.

[BEL90] J. R. Bellegarda and D. Nahamoo, "Tied Mixture Continuous Parameter Modeling for Speech Recognition", *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. 38, N°. 12, diciembre 1990, pp. 2033-2045.

[BEN90] Y. Bengio et al., "A Hybrid Coder for Hidden Markov Models Using a Recurrent Neural Network", *Proc. of ICASSP'90*, Alburquerque, Estados Unidos, 3-6 Abril 1990, pp. 537-540.

[BEN92] Y. Bengio et al., "Global Optimization of a Neural Network-Hidden Markov Model Hybrid", *IEEE Transactions on Neural Networks*, Vol. 3, N°. 2, marzo 1992, pp. 252- 259.

[BIT95] N. N. Bitar and C. Y. Espy-Wilson, "Speech Parametrization Based on Phonetic Features: Application to Speech Recognition", *Proc. of EUROSPEECH'95, Madrid, septiembre* 1995, pp. 1411-1414.

[BIT97] N. N. Bitar and C. Y. Espy-Wilson, "The Design of Acoustic Parameters for Speaker-Independent Speech Recognition", *Proc. of EUROSPEECH'97*, Rodas, Grecia, septiembre 1997, pp. 1239-1242.

[BLU79] S. E. Blumstein, "Acoustic invariance in speech production: Evidence from measurements of the spectral characteristics of stops consonants", *Journal of Acoustic Society of America*, Vol. 66, N° 4, octubre 1979, pp. 1001-1017.

[BOB98] J. Bobadilla, *Desarrollo de algoritmos basados en filtrado adaptativo y su aplicación en el estudio de la Fonética Acústica española*, Tesis Doctoral, Universidad Politécnica de Madrid, 1998.

[BON92] A. Bonneau et al., "A Model for Hypothetical Reasoning applied to Speech Recognition", *Proc. of the 10th European Conference on Artificial Intelligence, ECAI'92*, pp 553- 557.

[BON96] A. Bonneau, L. Djezzar and Y. Laprie, "Perception of the place of articulation of French stop bursts", *Journal of Acoustic Society of America*, Vol. 100, N° 1, julio 1996, pp. 555-564.

[BON97] A. Bonneau, "Relevant Spectral Information for the Identification of Vowel Features from Burst", *Proc. of EUROSPEECH'97*, Rodas, Grecia, septiembre 1997, pp. 2139-2142.

[BOO96] A. Boothroyd et al., "Effects of spectral smearing on phoneme and word recognition", *Journal of Acoustic Society of America*, Vol. 100, N° 3, septiembre 1996, pp. 1807-1818.

[BOU96] H. Boulard and H. Hermansky and N. Morgan, "Towards increasing speech recognition error rates", *Speech Communication*, Vol. 18, 1996, pp. 205-231.

[BUR96] D. Burshtein, "Robust Parametric Modeling of Durations in Hidden Markov Models", *IEEE Transactions on Speech and Audio Processing*, Vol. 4, N° 3, mayo 1996, pp. 240-242.

[BUS87] M. A. Bush and G. E. Kopec, "Network-Based Connected Digit Recognition", *IEEE Transactions on Acoustic, Speech and Signal Processing*, Vol. ASSP-35, N°. 10, octubre 1987, pp. 1401-1413.

[BUS95] P. A. Busby and G. L. Plant, "Formant frequency values of vowels produced by preadolescent boys and girls", *Journal of Acoustic Society of America*, Vol. 97, N° 4, abril 1995, pp. 2603-2606.

[CAR87] N. Carbonell, D. Fohr and J. P. Haton, "APHODEX, An Acoustic-Phonetic Decoding Expert System", *International Journal of Pattern Recognition and Artificial Intelligence*, Vol. 1, N°. 2, 1987, pp. 31-46.

[CAS95] A. Castellanos, J. M. Benedí and F. Casacuberta, "An Acoustic-Phonetic Analysis of Spanish Lombard Speech", *ESCA-NATO Tutorial and Research Workshop on Speech under Stress*, Lisboa, 14-15 septiembre, 1995, pp. 57-60.

[CER97] C. Cerisara and J. P. Haton, "Multi-Band Continuous Speech Recognition", *Proc. of EUROSPEECH'97*, Rodas, Grecia, septiembre 1997, pp. 1235-1238.

[CER98] C. Cerisara et al., "A Recombination Model for Multi-Band Speech Recognition", *Proc. of ICASSP'98*, Seattle, Estados Unidos, 12-15 mayo 1998, pp. 717-720.

[CHI97] J. T. Chien and H. C. Wang, "Telephone speech recognition based on Bayesian adaptation of hidden Markov models", *Speech Communication*, Vol. 22, 1997, pp. 369-384.

[CHO92] W. Chou, B. H. Juang and C. H. Lee, "Segmental GPD Training of HMM Based Speech Recognizer", *Proc. of ICASSP'92*, San Francisco, Estados Unidos, 23-26 marzo 1992, Vol. I, pp. 473-476.

[CLA97] T. Claes et al., "New Transformations of Cepstral Parameters for Automatic Vocal Tract Length Normalization in Speech Recognition", *Proc. of EUROSPEECH'97*, Rodas, Grecia, septiembre 1997, pp. 1363-1366.

[COH89] J. R. Cohen, "Application of an auditory model to speech recognition", *Journal of Acoustic Society of America*, Vol. 85, N° 6, junio 1989, pp. 2623-2629.

[COL96] R. A. Cole et al., "The Contribution of Consonants Versus Vowels to Word Recognition in Fluent Speech", *Proc. of ICASSP'96*, Atlanta, Estados Unidos, 7-10 mayo 1996, pp. 853-856.

[COL97] R. A. Cole et al., *Survey of he State of the Art in Human Language Technology*, Cambridge University Press, 1997.

[COO92] J. W. Cooley, "How the FFT Gained Acceptance", *IEEE SP Magazine*, enero 1992, pp. 10-13.

[DAS94] S. Das et al., "Adaptation Techniques for Ambience and Microphone compensation in the IBM Tangora Recognition System", *Proc. of ICASSP'94*, Adelaida, Australia, 19-22 abril 1994, Vol. I, pp.21-24.

[DAV52] K. H. Davis, R. Biddulph and S. Balashek, "Automatic Recognition of Spoken Digits", *Journal of Acoustic Society of America*, Vol. 24, N° 6, noviembre 1952, pp. 637-642.

[DAV80] S. B. Davis and P. Mermelstein, "Comparison of Parametric Representations for Monosillabic Word Recognition in Continuously Spoken Sentences", *IEEE Transactions on Acoustic, Speech, and Signal Processing*, Vol. ASSP-28, N° 4, agosto 1980, pp. 357-366.

[DEL93] J. R. Deller, J. G. Proakis and J. H. L. Hansen, Discrete-Time Processing of Speech Signals, Mac Millan, N. Y., 1993.

[DEL94] J. R. Deller, JR., "Tom, Dick and Mary Discover the DFT", *IEEE Signal Processing Magazine*, abril 1994, pp. 36-50.

[DEM84] R. De Mori and Y. F. Mong, "A System of Plans for Connected Speech Recognition", *Proc. of the National Conference on Artificial Intelligence AAAI'84*, Austin, Texas, Estados Unidos, 6-10 agosto 1984, pp. 92-95.

[DEN94a] L. Deng and D. X. Sun, "A statistical approach to automatic speech recognition using the atomic speech units constructed from overlapping articulatory features", *Journal of Acoustic Society of America*, Vol. 95, N° 5, mayo 1994, pp. 2702-2719.

[DEN94b] L. Deng and D. Sun, "Phonetic Classification and Recognition Using HMM Representation of Overlapping Articulatory Features for All Classes of English Sounds", *Proc. of ICASSP'94*, Adelaida, Australia, 19-22 abril 1994, Vol. I, pp. 45-48.

[DER86] A. M. Derouault and B. Merialdo, "Natural Language Modeling for Phoneme-to-Text Transcription", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. PAMI-8, N°. 6, noviembre 1986, pp. 742-749.

[DIA99] F. Díaz, *Utilización de métodos discriminantes en el modelado acústico de la voz*, Tesis Doctoral, Universidad Politécnica de Madrid, 1999.

[DIG94] V. Digalakis and H. Murveit, "Genones: Optimizing the Degree of Mixture Tying in a Large Vocabulary Hidden Markov Model Based Speech Recognizer", *Proc. of ICASSP'94*, Adelaida, Australia, 19-22 abril 1994, Vol. I, pp. 537-540.

[DJE95] L. Djezzar and J. P. Haton, "Exploiting Acoustic-Phonetic Knowledge and Neural Networks for Stop Recognition", *Proc. of EUROSPEECH'95, Madrid, septiembre* 1995, pp. 2217-2220.

[DOB97] S. Dobrisek, F. Mihelic and N. Pavesic, "A Multiresolutionally Oriented Approach for Determination of Cepstral Features in Speech Recognition", *Proc. of EUROSPEECH'97*, Rodas, Grecia, septiembre 1997, pp. 1367-1370.

[ELL62] L. L. Elliott, "Backward and Forward Masking of Probe Tones of Different Frequencies", *Journal of Acoustic Society of America*, Vol. 34, N° 8, agosto 1962, pp. 1116-1117.

[EPH87] Y. Ephraim, J. G. Wilpon and L. R. Rabiner, "A Linear Predictive Front-End Processor for Speech Recognition in Noisy Environments", *Proc. of ICASSP'87*, Dallas, Estados Unidos, 6-9 abril 1987, pp.1324-1327.

[EPH89] Y. Ephraim, A. Dembo and L. R. Rabiner, "A Minimum Discrimination Information Approach for Hidden Markov Modeling", *IEEE Transactions on Information Theory*, Vol. 35, N°. 5, septiembre 1989, pp. 1001-1013.

[ERL96] K. Erler and George H. Freeman, "An HMM-based speech recognizer using overlapping articulatory features", *Journal of Acoustic Society of America*, Vol. 100, N° 4, octubre 1996, pp. 2500-2513.

[ESP97] A. Esposito, "The amplitudes of the Peaks in the Spectrum: Data from [a] Context", *Proc. of EUROSPEECH'97*, Rodas, Grecia, septiembre 1997, pp. 1015-1018.

[ESS92] U. Essen and V. Steinbiss, "Cooccurrence Smoothing for Stochastic Language Modeling", *Proc. of ICASSP'92*, San Francisco, Estados Unidos, 23-26 marzo 1992, Vol. I, pp. 161-164.

[FAN73] G. Fant, *Speech Sounds and Features*, MIT Press, 1973.

[FIS98] A. Fischer and V. Stahl, "Subword Unit-Based Speech Recognition in Car Environments", *Proc. of ICASSP'98*, Seattle, Estados Unidos, 12-15 mayo 1998, pp. 257-260.

[FUR86] S. Furui, "Speaker-Independent Isolated Word Recognition Using Dynamic Features of Speech Spectrum", *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. ASSP-34, N°. 1, abril 1986, pp. 52-59.

[FUR89] S. Furui, *Digital Speech Processing, Synthesis and Recognition*, Marcel Dekker Inc., 1989.

[GAL92] M. J. F. Gales and S. Young, "An Improved Approach to the Hidden Markov Model Decomposition of Speech and Noise", *Proc. of ICASSP'92*, San Francisco, Estados Unidos, 23-26 marzo 1992, Vol. I, pp. 233-236.

[GAM97] S. Gamm, R. Haeb-Umbach and D. Langmann, "The development of a command-based speech interface for a telephone answering machine", *Speech Communication*, Vol. 23, 1997, pp. 161-171.

[GAR98] P. N. Garner and W. J. Holmes, "On the Robust Incorporation of Formant Features Into Hidden Markov Models for Automatic Speech Recognition", *Proc. of ICASSP'98*, Seattle, Estados Unidos, 12-15 mayo 1998, pp. 1-4.

[GAU94] J. L. Gauvain, "The LIMSI Continuous Speech Dictation System: Evaluation on the ARPA Wall Street Journal Task", *Proc. of ICASSP'94*, Adelaida, Australia, 19-22 abril 1994, Vol. I, pp. 557-560.

[GHI88] O. Ghitza, "Temporal non-place information in the auditory-nerve firing patterns as a front-end for speech recognition in a noisy environment", *Journal of Phonetics*, Vol. 16, 1988, pp. 109-123.

[GOD92] D. Goddeau and V. Zue, "Integrating Probabilistic LR Parsing Into Speech Understanding Systems", *Proc. of ICASSP'92*, San Francisco, Estados Unidos, 23-26 marzo 1992, Vol. I, pp. 181-184.

[GON88] Y. F. Gong and J. P. Haton, "A Specialist Society for Continuous Speech Understanding", *Proc. of ICASSP'88*, Nueva York, Estados Unidos, 11-14 abril 1988, pp. 627-630.

[GON94] Y. Gong and J. P. Haton, "Stochastic Trajectory Modeling for Speech Recognition", *Proc. of ICASSP'94*, Adelaida, Australia, 19-22 abril 1994, Vol. I, pp. 57-60.

[GOO53] I. J. Good, "The Population Frequencies of Species and the Estimation of Population Parameters", *Biometrika*, Vol. 40, Partes 3 y 4, diciembre 1953, pp. 237-264.

[GOP89] P. S. Gopalakrisnan et al., "A Generalization of the Baum Algorithm to Rational Objective Functions", *Proc. of ICASSP'89*, Glasgow, Reino Unido, 23-26 mayo 1989, pp. 631-634.

[GOP95a] P. S. Gopalakrishnan, "Continuous Speech Recognition", *Modern Methods of Speech Processing*, R. P. Ramachandran and R, J. Mammone editores, Kluwer Academic Publishers, 1995, pp. 185-212.

[GOP95b] P. S. Gopalakrisnan, L. R. Bahl and R. L. Mercer, "A Tree Search Strategy for Large-Vocabulary Continuous Speech Recognition", *Proc. of ICASSP'95*, Detroit, Estados Unidos, 9-12 mayo 1995, Detroit, Estados Unidos, 9-12 mayo 1995, pp. 572-575.

[GOU97] E. B. Gouvêa and R. Stern, "Speaker Normalization Through Formant-Based Warping of the Frequency Scale", *Proc. of EUROSPEECH'97*, Rodas, Grecia, septiembre 1997, pp. 1139-1142.

[GRA84] R. M. Gray, "Vector Quantization", *IEEE ASSP Magazine*, abril 1984, pp. 4-29.

[GRE86] P. D. Green and A. R. Wood, "A Representational Approach to Knowledge-Based Acoustic-Phonetic Processing in Speech Recognition", *Proc. of ICASSP'86*, pp. 1205-1208.

[GUP87] V. N. Gupta, M. Lennig and P. Mermesltein, "Integration of Acoustic Information in a Large Vocabulary Word Recognizer", *Proc. of ICASSP'87*, Dallas, Estados Unidos, 6-9 abril 1987, pp. 697-700.

[GUP88] V. N. Gupta, M. Lennig and P. Mermelstein, "Fast search strategy in a large vocabulary word recognizer", *Journal of Acoustic Society of America*, Vol. 44, N° 6, diciembre 1988, pp. 2007-2017.

[GUP95] V. Gupta and M. Lenning, "Large Vocabulary Isolated Word Recognition", *Modern Methods of Speech Processing*, R. P. Ramachandran and R, J. Mammone editores, Kluwer Academic Publishers, 1995, pp. 213-230.

[HAE92] R. Haeb-Umbach and H. Ney, "Linear Discriminant Analysis for Improved Large Vocabulary Continuous Speech Recognition", *Proc. of ICASSP'92*, San Francisco, Estados Unidos, 23-26 marzo 1992, Vol. I, pp. 13-16.

[HAL97] A. K. Halberstadt and J. R. Glass, "Heterogeneous Acoustic Measurements for Phonetic Classification", *Proc. of EUROSPEECH'97*, Rodas, Grecia, septiembre 1997, pp. 401-404.

[HAN90] B. A. Hanson and T. H. Applebaum, "Robust Speaker-Independent Word Recognition Using Static, Dynamic and Acceleration Features: Experiments with Lombard and Noisy Speech". *Proc. of ICASSP'90*, Alburquerque, Estados Unidos, 3-6 abril 1990, pp. 857-860.

[HAN95] J. H. L. Hansen, "Analysis and Compensation of Speech Under Stress & Noise for Environmental Robustness in Speech Recognition", *ESCA-NATO Tutorial and Research Workshop on Speech under Stress*, Lisboa, 14-15 septiembre, 1995, pp. 91-98.

[HAN96] J. H. L. Hansen and B. Womack, "Feature Analysis and Neural Network-Based Classification of Speech Under Stress", *IEEE Transactions on Speech and Audio Processing*, Vol. 4, N° 4, julio 1996, pp. 307-313.

[HAT85] J. P. Haton, "Knowledge-Based and Expert Systems in Automatic Speech Recognition", *New Systems and Architectures for Automatic Speech Recognition and Synthesis*, R. De Mori and C. Y. Suen editores, Springer-Verlag, Berlin, 1985.

[HAY96] S. Haykin, *Adaptive Filter Theory*, 2nd Ed., Prentice Hall, Englewood Cliffs, N. J., 1996

[HEC68] M. H. L. Hecker et al., "Manifestations of Task-Induced Stress in the Acoustic Speech Signal", *Journal of Acoustic Society of America*, Vol. 44, N° 4, abril 1968, pp. 993-1001.

[HEN92] C. Henton, "Acoustic variability in the vowels of female and male speakers", *Journal of Acoustic Society of America*, Vol. 91, N° 4, abril 1992, pp. 2387.

[HER90] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech", *Journal of Acoustic Society of America*, Vol. 87, N° 4, abril 1990, pp. 1738-1752.

[HER92] H. Hermansky et al., "RASTA-PLP speech analysis technique", *Proc. of ICASSP'92*, San Francisco, Estados Unidos, 23-26 marzo 1992, pp. 121-124.

[HER93] H. Hermansky, N. Morgan and H. G. Hirsch, "Recognition of Speech in Additive and Convolutional Noise Based on RASTA Spectral Processing", *Proc. of ICASSP'93*, Minneapolis, Estados Unidos, 27-30 abril 1993, Vol. II, pp. 83-86.

[HER94] H. Hermansky and N. Morgan, "RASTA Processing of Speech", *IEEE Transactions on Speech and Audio Processing*, Vol. 2, N°. 4, octubre 1994, pp. 578-589.

[HIL95] J. Hillenbrand et al., "Acoustic characteristics of American English vowels", *Journal of Acoustic Society of America*, Vol. 97, N° 5, mayo 1995, pp. 3099-3111.

[HOL86] J. N. Holmes, "Normalization in Vowel Perception", *Invariance and Variability in Speech Processes*, J. S. Perkell and D. H. Klatt editores, Lawrence Erlbaum Associates Publishers, 1986, pp. 346- 359.

[HOL97] J. N. Holmes, W. J. Holmes and P. N. Garner, "Using Formant Frequencies in Speech Recognition", *Proc. of EUROSPEECH'97*, Rodas, Grecia, septiembre 1997, pp. 2083-2086.

[HON91] H. W. Hon and K. F. Lee, "CMU Robust Vocabulary-Independent Speech Recognition System", *Proc. of ICASSP'91*, Toronto, Canadá, 14-17 mayo 1991, pp. 889-892.

[HON92] H. W. Hon and K. F. Lee, "Vocabulary Learning and Environment Normalization in Vocabulary-Independent Speech Recognition", *Proc. of ICASSP'92*, San Francisco, Estados Unidos, 23-26 marzo 1992, Vol. I, pp. 485-488.

[HOU71] T. Houtgast, "Psychophysical Evidence for Lateral Inhibition in Hearing", *Journal of Acoustic Society of America*, Vol. 51, N° 6, Parte 2, abril 1971, pp. 1885-1894.

[HUA90] X. D. Huang, Y. Ariki and M. A. Jack, *Hidden Markov Models for Speech Recognition*, Edinburgh University Press, 1990.

[HUA93] X. Huang and K. F. Lee, "On Speaker-Independent, Speaker-Dependent and Speaker-Adaptive Speech Recognition", *IEEE Transactions on Speech and Audio Proc.*, Vol. 1, N° 2, abril 1993, pp. 150-157.

[HUN89] M. J. Hunt and C. Lefèvre, "A Comparison of Several Acoustic Representations for Speech Recognition with Degraded and Undegraded Speech", *Proc. of ICASSP'89*, Glasgow, Reino Unido, 23-26 mayo 1989, pp. 262-265.

[ITA70] F. Itakura and S. Saito, "A Statical Method for Estimation of Speech Spectral Density and Formant Frequencies", *Electron. Communication*, Vol. 53, pp. 36-43

[IVO.html] http://www.tamarisco.fi.upm.es/projects//IVORY

[JAK63] R. Jakobson, G. Fant and M. Halle, *Preliminaries to Speech Analysis. The Distitictive Features and their Correlates*, MIT Press, 1963.

[JAR96] M. Jardino, "Multilingual Stochastic N-Gram Class Language Models", *Proc. of ICASSP'96*, Atlanta, Estados Unidos, 7-10 mayo 1996, pp. 161-163.

[JAV83] E. Javel et al., "Suppression of auditory nerve responses. II. Suppression threshold and growth, iso-suppression contours", *Journal of Acoustic Society of America*, Vol. 74, N° 3, septiembre 1983, pp. 801-813.

[JEL75] F. Jelinek, L. R. Bahl and R. L. Mercer, "Design of a Linguistic Statistical Decoder for the Recognition of the Continuous Speech", *IEEE Transactions on Information Theory*, Vol. IT-21, N°. 3, mayo 1975, pp. 250- 256.

[JEL80] F. Jelinek and R. L. Mercer, "Interpolated Estimation of Markov Source Parameters from Sparse Data", *Pattern Recognition in Practice*, E. S. Gelsema and L. N. Kanal editores, North-Holland Publising Company 1980.

[JEL85] F. Jelinek, "The Development of an Experimental Discrete Dictation Recognizer", *Proceedings of the IEEE*, Vol. 73, N°. 11, noviembre 1985, pp. 1616-1623.

[JEL92] F. Jelinek, R. L. Mercer and S. Roukos, "Principles of Lexical Language Modeling for Speech Recognition", *Advances in Speech Signal Processing*, S. Furui and M. M. Sondhi editores, Marcel Dekker Inc., 1992, pp. 651-699.

[JEL94] F. Jelinek, "Training and Search Methods for Speech Recognition", *Voice Communication Between Humans ans Machines*, D. B. Roe and J. G. Wilpon editores, National Academy Press, Washington D. C., 1994, pp. 199-214.

[JEL96] F. Jelinek, "Five speculations (and a divertimento) on the themes of H. Boulard, H. Hermansky and N. Morgan", *Speech Communication*, Vol. 18, 1996, pp. 242-246.

[JEL97] F. Jelinek, *Statistical methods for Speech Recognition*, MIT Press, 1997.

[JOH97] K. Johnson, *Acoustic and Auditory Phonetics*, Blackwell Publishers, Cambridge, Massachussetts, 1997.

[JUA92] B. H. Juang and S. Kataragi, "Discriminative Learning for Minimum Error Classification", *IEEE Transactions on Signal Processing*, Vol. 40, N°. 12, diciembre 1992, pp. 3043-3054.

[JUA98] B. H. Juang, "The Past, Present, and Future of Speech Processing", *IEEE Signal Processing Magazine*, mayo 1998, pp. 24-48.

[JUN89] J. C. Junqua and H. Wakita, "A Comparative Study of Cepstral Lifters and Distance Measures for All Pole Models of Speech in Noise", *Proc. of ICASSP'89*, Glasgow, Reino Unido, 23-26 mayo 1989, pp. 476-479.

[JUN93] J. C. Junqua, "The Lombard reflex and its role on human listeners and automatic speech recognizers", *Journal of Acoustic Society of America*, Vol. 93, N° 1, enero 1993, pp. 510-524.

[JUN94] J. C. Junqua, B. Mak and B. Reaves, "A Robust Algorithm for Word Boundary Detection in the Presence of Noise", *IEEE Transactions on Speech and Audio Processing*, Vol. 2, N°. 3, julio 1994, pp. 406-412.

[JUN95] C. Junqua, "The influence of Acoustics on Speech Production: a Noise-Induced Stress Phenomenon Known as the Lombard Reflex", *ESCA-NATO Tutorial and Research Workshop on Speech under Stress*, Lisboa, 14-15 septiembre 1995, pp. 83-90.

[JUN96] J. C. Junqua and J. P. Haton, *Robustness in Automatic Speech Recognition*, Kluwer Academic Publishers, 1996.

[JUN97] J. C. Junqua, "SmarTspelT™: A Multipass Recognition System for Name Retrieval over the Telephone", *IEEE Transactions on Speech and Audio Processing*, Vol. 5, N° 2, marzo 1997, pp. 173-182.

[KAS95] K. Kasper et al., "A Fully Recurrent Neural Network for Recognition of Noisy Telephone Speech", *Proc. of ICASSP'95*, Detroit, Estados Unidos, 9-12 mayo 1995, Detroit, Estados Unidos, 9-12 mayo 1995, pp. 3331-3334.

[KAT87] S. M. Katz, "Estimation of Probabilities from Sparse Data for the Language Model Component of a Speech Recognizer", *IEEE Transactions on Acoustic, Speech and Signal Processing*, Vol. ASSP-35, N°. 3, marzo 1987, pp. 400-401.

[KEW96] D. Kewley-Port et al., "Fundamental frequency effects on thresholds for vowel formant discrimination", *Journal of Acoustic Society of America*, Vol. 100, N° 4, octubre 1996, pp. 2462-2470.

[KLA90] D. H. Klatt and L. C. Klatt, "Analysis, synthesis, and perception of voice quality variations among female and male talkers", *Journal of Acoustic Society of America*, Vol. 87, N° 2, febrero 1990, pp. 820-857.

[KOE94] J. Koehler et al., "Integrating RASTA-PLP Into Speech Recognition", *Proc. of ICASSP'87*, Dallas, Estados Unidos, 6-9 abril 1987, Vol. I, pp. 421-424.

[KOH90] T. Kohonen, "The Self-Organizing Map", *Proceedings of the IEEE*, Vol. 78, N° 9, pp. 1464-1480, septiembre 1990.

[KOM97] R. Kompe, *Prosody in Speech Understanding Systems*, Springer-Verlag, Berlin, 1997.

[KRA94] P. Kraniauskas, "A Plain Man's Guide to the FFT", *IEEE Signal Processing Magazine*, abril 1994, pp. 24-35.

[KUB94] F. Kubala et al., "Comparative Experiments on Large Vocabulary Speech Recognition", *Proc. of ICASSP'94*, Adelaida, Australia, 19-22 abril 1994, Vol. I, pp. 561-564.

[KUH94] R. Kuhn, R. De Mori and E. Millien, "Learning Consistent Semantics from Training Data" *Proc. of ICASSP'94*, Adelaida, Australia, 19-22 abril 1994, Vol. II, pp. 237-240.

[KUW97] H. Kuwabara, "Acoustic and Perceptual Properties of Phonemes in Continuous Speech as a Function of Speaking Rate", *Proc. of EUROSPEECH'97*, Rodas, Grecia, septiembre 1997, pp. 1003-1006.

[LAB86] W. Labov, "Sources of Inherent Variation in the Speech Process", *Invariance and Variability in Speech Processes*, J. S. Perkell and D. H. Klatt editores, Lawrence Erlbaum Associates Publishers, 1986, pp. 402-425.

[LAD96] P. Ladefoged, *Elements of Acoustic Phonetics*, 2nd ed., The University of Chicago Press, Chicago, 1996.

[LAM81] L. F. Lamel et al., "An Improved Endpont Detector for Isolated Word Recognition", *IEEE Transactions on Acoustic, Speech and Signal Processing*, Vol. ASSP-29, N°. 4, agosto 1981, pp. 777-785.

[LAP94] Y. Laprie and M. O. Berger, "A new paradigm for reliable automatic formant tracking", *Proc. of ICASSP'94*, Adelaida, Australia, 19-22 abril 1994, pp. 201-204.

[LAP96] P. Lapsley and G. Blalock, "How to estimate DSP processor performance", *IEEE Spectrum*, julio 1996, pp. 74-78

[LAU93] R. Lau, R. Rosenfeld and S. Roukos, "Trigger-Based Language Models: A Maximum Entropy Approach", *Proc. of ICASSP'93*, Minneapolis, Estados Unidos, 27-30 abril 1993, Vol. II, pp. 45-48.

[LEC89] Y. Lecomte et al., "Car Noise Processing for Speech Input", *Proc. of ICASSP'89*, Glasgow, Reino Unido, 23-26 mayo 1989, pp. 512-515.

[LEC94] P. Le Cerf, W. Ma and D. V. Compernolle, "Multilayer Perceptrons as Labelers for Hidden Markov Models", *IEEE Transactions on Speech and Audio Processing*, Vol. 2, N°. 1, Parte II, enero 1994, pp. 185-193.

[LEE88a] C. H. Lee, F. K. Soong and B. H. Juang, "A Segment Model Based Approach to Speech Recognition", *Proc. of ICASSP'88*, Nueva York, Estados Unidos, 11-14 abril 1988, pp. 501-504.

[LEE88b] E. A. Lee, "Programmable DSP Architectures: Part I", *IEEE ASSP Magazine*, octubre 1988, pp. 4-19.

[LEE89a] K. F. Lee et al., "The SPHINX Speech Recognition System", *Proc. of ICASSP'89*, Glasgow, Reino Unido, 23-26 mayo 1989, pp. 445-448.

[LEE89b] K. F. Lee, *Automatic Speech Recognition: The Development of the SPHINX System*, Kluwer Academic Publishers, Boston, Massachussetts, 1989.

[LEE89c] K. F. Lee and H. W. Hon, "Speaker-Independent Phone Recognition using Hidden Markov Models", *IEEE Transactions on Acoustic, Speech, and Signal Processing*, Vol. ASSP-37, N° 11, febrero 1996, pp. 230-239.

[LEE90a] K. F. Lee, "Context-Dependent Phonetic Hidden Markov Models for Speaker-Independent Continuous Speech Recognition", *IEEE Transactions on Acoustics, Speech and Signal Processing*, Vol. 38, N°. 4, abril 1990, pp. 599-609.

[LEE90b] K. F. Lee, H. W. Hon and R. Reddy, "An Overview of the SPHINNX Speech Recognition System", *IEEE Transactions on Acoustics, Speech and Signal Processing*, Vol. 38, N°. 1, enero 1990, pp. 35-45.

[LEE90c] K. F. Lee et al., "Allophone Clustering for Continuous Speech Recognition", *Proc. of ICASSP'90*, Alburquerque, Estados Unidos, 3-6 Abril 1990, pp. 749-752.

[LEE91] C. H. Lee, C. H. Lin and B. H. Juang, "A Study on Speaker Adaptation of the Parameters of Continuous Density Hidden Markov Models", *IEEE Transactions on Signal Processing*, Vol. 39, N°. 4, abril 1991, pp. 806-814.

[LEE95] T. Lee, P. C. Ching and L. W. Chan, "An RNN Based Speech Recognition System with Discriminative Training", *Proc. of EUROSPEECH'95, Madrid, septiembre* 1995, pp. 1667-1670.

[LEE96] M. R. Leek and V. Summers, "Reduced frequency selectivity and the preservation of spectral contrast in noise", *Journal of Acoustic Society of America*, Vol. 100, N° 3, septiembre 1996, pp. 1796-1806.

[LEE98] L. Lee and R. Rose, "A Frequency Warping Approach to Speaker Normalization", *IEEE Transactions on Speech and Audio Processing*, Vol. 6, N° 1, enero 1998, pp. 49-60.

[LIN80] Y. Linde, A. Buzo and R. M. Gray, "An Algorithm for Vector Quantizer Design", *IEEE Transactions on Communications*, Vol. 28, N°. 1, enero 1980, pp. 84- 95.

[LIN97] M. Lincoln, S. Cox and S. Ringland, "A Fast Method of Speaker Normalization Using Formant Estimation", *Proc. of EUROSPEECH'97*, Rodas, Grecia, septiembre 1997, pp. 2095-2098.

[LIP82] L. A. Liporace, "Maximum Likehood Estimation for Multivariate Observations of Markov Sources", *IEEE Transactions on Information Theory*, Vol. IT-28, N°. 5, septiembre 1982, pp. 729-734.

[LIP87a] R. P. Lippmann, "An Introduction to Computing with Neural Nets", *IEEE ASSP Magazine*, pp. 4-23, abril 1987.

[LIP87b] R. P. Lippmann, E. A. Martin and D. B. Paul, "Multi-style Trainning for Robust isolated-word speech Recognition", *Proc. of ICASSP'87*, Dallas, 6-9 abril 1987, pp. 705-708.

[LIU94] F. H. Liu et al., "Environment Normalization for Robust Speech Recognition Using Direct Cepstral Comparison", *Proc. of ICASSP'94*, Adelaida, Australia, 19-22 abril 1994, Vol. II, pp. 61-64.

[LIU96] S. A. Liu, "Landmark detection for distinctive feature-based speech recognition", *Journal of Acoustic Society of America*, Vol. 100, N° 5, noviembre 1996, pp. 3417-3430.

[LOC92] P. Lockwood, J. Boudy and M. Blanchet, "Non-Linear Spectral Subtraction (NSS) and Hidden Markov Models for Robust Speech Recognition in Car Noise Environments", *Proc. of ICASSP'92*, San Francisco, Estados Unidos, 23-26 marzo 1992, Vol. I, pp. 265-268.

[LYO82] R. F. Lyon, "A Computational Model of Filtering, Detection, and Compression in the Cochlea", *Proc. of ICASSP'82*, Paris, Francia, 3-5 mayo 1982, pp. 1282-1285.

[MAK75] J. Makhoul, "Linear Prediction: A tutorial Review", *Proc. of the IEEE*, Vol. 63, abril 1975, pp. 124-143.

[MAK85] J. Makhoul, S. Roucos and H. Gish, "Vector Quantization in Speech Coding", *Proc. of the IEEE*, Vol. 73, N° 11, noviembre 1985, pp. 1551-1588.

[MAL92] G. Maltese and F. Mancini, "An Automatic Technique to Include Grammatical and Morphological Informatical and Morphological Information in a Trigram-Based Statistical Language Model", *Proc. of ICASSP'92*, San Francisco, Estados Unidos, 23-26 marzo 1992, Vol. I, pp. 157-160.

[MAR72] J. D. Markel, "Digital Inverse Filtering- A new Tool for Formant Trajectory Estimation", *IEEE Transactions Audio Electroacoust.*, Vol. 20, pp. 129-137.

[MAR94] J. F. Mari and J. P. Haton, "Automatic Word Recognition based on Second-Order Hidden Markov Models", *ICSLP'94*, Yokohama, Japón, septiembre 1994, pp. 247-250.

[MAR97] J. F. Mari and J. P. Haton, "Automatic Word Recognition Based on Second-Order Hidden Markov Models", *IEEE Transactions on Speech and Audio Processing*, Vol. 5, N° 1, enero 1997, pp. 22-25.

[MAR98] F. Martínez, D. Tapias and J. Álvarez, "Towards Speeech Rate Independence in Large Vocabulary Continuous Speech Recognition", *Proc. of ICASSP'98*, Seattle, Estados Unidos, 12-15 mayo 1998, Seattle, Estados Unidos, 12-15 mayo 1998, pp. 725-728.

[MCC74] S. S. McCandless, "An Algorithm for Automatic Formant Extraction Using Linear Prediction Spectra", *IEEE Transactions on Acoustic, Speech, and Signal Processing*, Vol. ASSP-22, N° 2, abril 1974, pp. 135-141.

[MCC98] P. McCourt, S. Vaseghi and N. Harte, "Multi_Resolution Cepstral Features for Phoneme Recognition Across Speech Sub-Bands" *Proc. of ICASSP'98*, Seattle, Estados Unidos, 12-15 mayo 1998, pp. 557-560.

[MER75] P. Mermelstein, "Automatic segmentation of speech into syllabic units", *Journal of Acoustic Society of America*, Vol. 58, N° 4, octubre 1975, pp. 880-883.

[MER91] N. Merhav and Y. Ephraim, "Maximum Likehood Hidden Markov Modeling Using a Dominant Sequence of States", *IEEE Transactions on Signal Processing*, Vol. 39, N°. 9, septiembre 1991, pp. 2111-2114.

[MIL55] G. A. Miller and P. E. Nicely, "An Analysis of Perceptual Confusions Among Some English Consonants", *Journal of Acoustic Society of America*, Vol. 27, N° 2, marzo 1955, pp. 338-352.

[MIL89] J. D. Miller, "Auditory-perceptual interpretation of the vowel", *Journal of Acoustic Society of America*, Vol. 85, N° 5, mayo 1989, pp. 2114-2134.

[MIL97] B. Milner, "Cepstral-Time Matrices and LDA for Improved Connected Digit and Sub-Word Recognition Accuracy", *Proc. of EUROSPEECH'97*, Rodas, Grecia, septiembre 1997, pp. 405-408.

[MOR94] P. J. Moreno and R. M. Stern, "Sources of Degradation of Speech Recognition in the Telephone Network" *Proc. of ICASSP'94*, Adelaida, Australia, 19-22 abril 1994, Vol. I, pp. 109-112.

[MOR95] P. J. Moreno, B. Raj and R. Stern, "A Unified Approach for Robust Speech Recognition", *Proc. of EUROSPEECH'95, Madrid, septiembre* 1995, pp. 481-484.

[MOR95] N. Morgan and H. Boulard, "Continuous Speech Recognition. An Introduction to the hybrid HMM/Connectionist approach", *IEEE Signal Processing Magazine*, Vol. 12, N° 3, pp. 25-42, mayo 1995.

[MOR97] A. C. Morris et al., "Human and Machine Identification of Consonantal Place of Articulation from Vocalic Transition Segments", *Proc. of EUROSPEECH'97*, Rodas, Grecia, septiembre 1997, pp. 2123-2126.

[MUR93] H. Murveit et al., "Large-Vocabulary Dictation Using SRI's Decipher™ Speech Recognition System: Progressive Search Techniques", *Proc. of ICASSP'93*, Minneapolis, Estados Unidos, 27-30 abril 1993, Vol. II, pp. 319-322.

[NAB96] A. K. Nábêlek et al., "Cues for perception of synthetic and natural diphthongs in either noise or reverberation", *Journal of Acoustic Society of America*, Vol. 99, N° 3, marzo 1996, pp. 1742-1753.

[NEA97] A. Neagu and G. Bailly, "Relative Contributions of Noise Burst and Vocalic Transitions to the Perceptual Identification of Stop Consonants", *Proc. of EUROSPEECH'97*, Rodas, Grecia, septiembre 1997, pp. 2175-2178.

[NEU94] L. Neumeyer and M. Weintraub, "Probabilistic Optimum Filtering for Robust Speech Recognition", *Proc. of ICASSP'94*, Adelaida, Australia, 19-22 abril 1994, Vol. I, pp. 417-420.

[NEY91] H. Ney and U. Essen, "On Smoothing Techniques for Bigram-Based Natural Language Modelling", *Proc. of ICASSP'91*, Toronto, Canadá, 14-17 mayo 1991, pp. 825-828.

[NEY92a] H. Ney, D. Mergel and A. Noll, "Data Driven Search Organization for Continuous Speech Recognition", *IEEE Transactions on Signal Processing*, Vol. 40, N°. 2, febrero 1992, pp. 272-281.

[NEY92b] H. Ney et al, "Improvements in Beam Search for 10000-Word Continuous Speech Recognition", *Proc. of ICASSP'92*, San Francisco, Estados Unidos, 23-26 marzo 1992, Vol. I, pp. 9-12.

[NIC97] S. Nicholson, B. Milner and S. Cox, "Evaluating Feature Set Performance Using the F-Ratio and J-Measures", *Proc. of EUROSPEECH'97*, Rodas, Grecia, septiembre 1997, pp. 413-416.

[NIL80] N. Nilsson, *Principles of Artificial Intelligence*, Palo Alto, California, Tioga, 1980.

[NIR94] M. Niranjan and Y. J. Cox, "Recursive Tracking of Formants in Speech Signals", *Proc. of ICASSP'94*, Adelaida, Australia, 19-22 abril 1994, pp. 205-208.

[NIT98] T. Nitta, "A Novel Feature-Extraction for Speech Recognition Based on Multiple Acoustic-Feature Planes", *Proc. of ICASSP'98*, Seattle, Estados Unidos, 12-15 mayo 1998, pp. 29-32.

[NIY98] P. Niyogi and P. Ramesh, "Incorporating Voice Onset Time to Improve Letter Recognition Accuracies", *Proc. of ICASSP'98*, Seattle, Estados Unidos, 12-15 mayo 1998, pp. 13-16.

[NOR91] Y. Normandin and S. D. Morgera, "An Improved MMIE Training Algorithm for Speaker-Independent, Small Vocabulary, Continuous Speech Recognition", *Proc. of ICASSP'91*, Toronto, Canadá, 14-17 mayo 1991, pp. 537-540.

[OH95] S. Oh and V. Viswanathan, "Microphone Array for Hands-Free Voice Communication in a Car", *Modern Methods of Speech Processing*, R. P. Ramachandran and R, J. Mammone editores, Kluwer Academic Publishers, 1995, pp. 351-375.

[OHA95] J. J. Ohala, "Clear Speech Does Not Exaggerate Phonemic Contrast", *Proc. of EUROSPEECH'95, Madrid, septiembre* 1995, pp. 1323-1326.

[OHD96] R. N. Ohde and M. T. Ochs, "The effect of segment duration on the perceptual integration of nasals for adult and children speech", *Journal of Acoustic Society of America*, Vol. 100, N° 4, octubre 1996, pp. 2486-2499.

[OKA98] S. Okawa, E. Bocchieri and A. Potamianos, "Multi-Band Speech Recognition in Noisy Environments", *Proc. of ICASSP'98*, Seattle, Estados Unidos, 12-15 mayo 1998, pp. 641-644.

[OPP75] A. V. Oppenheim and R. W. Schafer, *Digital Signal Processing*, Prentice Hall, Englewood Cliffs, N. J., 1975.

[OSH92] D. O'Shaughnessy, "Recognition of Hesitations in Spontaneous Speech", *Proc. of ICASSP'92*, San Francisco, Estados Unidos, 23-26 marzo 1992, Vol. I, pp. 521-524.

[OST96] M. Ostendorf, V. V. Digalakis and O. A. Kimball, "From HMM's to Segments Models: A Unified View of Stochastic Modeling for Speech Recognition", *IEEE Transactions on Speech and Audio Processing*, Vol. 4, N° 5, septiembre 1996, pp. 360-378.

[PAL93] K. K. Paliwal, "Use of Temporal Correlation Between Successive Frames in a Hidden Markov Model Based Speech Recognizer", *Proc. of ICASSP'93*, Minneapolis, Estados Unidos, 27-30 abril 1993, Vol. II, pp.215-218.

[PAL98] K. K. Paliwal, "Spectral Subband Centroid Features for Speech Recognition", *Proc. of ICASSP'98*, Seattle, Estados Unidos, 12-15 mayo 1998, pp. 617-620.

[PAU91] D. B. Paul, "Algorithms for An Optimal A* Search and Linearizing the Search in the Stack Decoder", *Proc. of ICASSP'91*, Toronto, Canadá, 14-17 mayo 1991, pp. 693-696.

[PAU92] D. B. Paul, "An Efficient A* Stach Decoder Algorithm for Continuous Speech Recognition with a Stochastic Language Model", *Proc. of ICASSP'92*, San Francisco, Estados Unidos, 23-26 marzo 1992, Vol. I, pp. 25-28.

[PEI95] A. M. Peinado et al., "MCE Estimation of VQ Parameters for MVQHMM Speech Recognition", *Proc. of EUROSPEECH'95, Madrid, septiembre* 1995, pp. 533-536.

[PEI96] A. M. Peinado et al., "Discriminative Codebook Design Using Multiple Vector Quantization in HMM-Based Speech Recognizers", *IEEE Transactions on Speech and Audio Processing*, Vol. 4, N° 2, marzo 1996, pp. 89-95.

[PET52] G. E. Peterson and H. L. Barney, "Control Methods Used in a Study of the Vowels", *Journal of Acoustic Society of America*, Vol. 246, N° 2, marzo 1952, pp. 175-184.

[PFA98] T. Pfau and G. Ruske, "Estimating the Speaking Rate by Vowel Detection", *Proc. of ICASSP'98*, Seattle, Estados Unidos, 12-15 mayo 1998, pp. 945-948.

[PIC57] J. M. Pickett, "Perception of Vowels Heard in Noises of Various Spectra", *Journal of Acoustic Society of America*, Vol. 29, N° 5, mayo 1957, pp. 613-620.

[PIC90] J. Picone, "Continuous Speech Recognition Using Markov Models", *IEEE ASSP Magazine*, Vol. 7, N° 3, julio 1990, pp. 26-41.

[PIC93] J. W. Picone, "Signal Modeling Techniques in Speech Recognition", Proc. of the IEEE, Vol. 81, N° 9, septiembre 1993, pp. 1215-1247.

[PIE69] J. R. Pierce, "Whither Speech Recognition?", *Journal of Acoustic Society of America*, Vol. 46, N° 4, Parte 2, junio 1969, pp. 1049-1050.

[PIS97] D. B. Pisoni, "Some Thoughts on 'Normalization' in Speech Perception", *Talker Variability in Speech Processing*, K. Johnson and J. W. Mullenix editores, Academic Press Limited, 1997, pp. 9-32.

[PLA95] F. Plante and W. A. Ainsworth, "Formant tracking using reasigned spectrum", *Proc. of EUROSPEECH'95*, Madrid, septiembre 1995, pp. 741-744.

[POR81] R. F. Port, "Linguistic timing factors in combination", *Journal of Acoustic Society of America*, Vol. 69, N° 1, enero 1981, pp. 262-274.

[POR84] J. E. Porter and S. F. Boll, "Optimal Estimators for Spectral Restoration of Noisy Speech", *Proc. of ICASSP'84*, San Diego, Estados Unidos, 19-21 marzo 1984, pp. 18A.2.1-18A.2.4.

[POR88] A. B. Poritz, "Hidden Markov Models: A Guided Tour", *Proc. of ICASSP'88*, Nueva York, Estados Unidos, 11-14 abril 1988, pp. 7-13.

[QUI88] A. Quilis, *Fonética Acústica de la lengua española*, Editorial Gredos, Madrid, 1993.

[QUI93] A. Quilis, *Tratado de Fonología y Fonética españolas*, Editorial Gredos, Madrid, 1993.

[QUI96] A. Quilis, *Curso de Fonética y Fonología españolas*, Consejo Superior de Investigaciones Científicas, Madrid, 1996.

[RAB81] L. R. Rabiner and S. E. Levinson, "Isolated and Connected Word Recognition- Theory and Selected Applications", *IEEE Transactions on Communications*, Vol. COM-29, N°. 5, mayo 1981, pp. 621-659.

[RAB86] L. R. Rabiner and B. H. Juang, "An Introduction to Hidden Markov Models", *IEEE ASSP Magazine*, Vol. 3, N° 1, enero 1996, pp. 4-16

[RAB89] L. R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition", *Proc. of the IEEE*, Vol. 77, N° 2, febrero 1989, pp. 257-286

[RAB93] L. R. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*, Prentice-Hall, Englewood Cliffs, N. J., 1993.

[RAB96] L. R. Rabiner, B. H. Juang and C. H. Lee, "An Overview of Automatic Speech Recognition", *Automatic Speech and Speaker Recognition: Advanced Topics*, C. H. Lee, F. K. Soong and K. K. Paliwal editores, Kluwer Academic Publisher, 1996, pp. 1-30.

[RAH97] M. Rahim, Y. Bengio and Y. LeCun, "Discriminative Features and Model Design for Automatic Speech Recognition", *Proc. of EUROSPEECH'97*, Rodas, Grecia, septiembre 1997, pp. 75-78.

[RAJ86] P. K. Rajasekaran, G. R. Dodington and J. W. Picone, "Recognition of Speech Under Stress and in Noise", *Proc. of ICASSP'86*, pp. 733-736.

[REI95] W. Reich and G. Ruske, "Discriminative Training for continuous speech recognition", *Proc. of EUROSPEECH'95*, Madrid, septiembre 1995, pp. 537-540.

[REM91] R. E. Remez, "On the perceptual differentiation of spontaneous and prepared speech", *Journal of Acoustic Society of America*, Vol. 89, N° 4, Parte 2, abril 1991, pp. 2011-2012.

[RIG94] G. Rigoll, "Maximum Mutual Information Neural Networks for Hybrid Connectionist-HMM Speech Recognition Systems", *IEEE Transactions on Speech and Audio Processing*, Vol. 2, N°. 1, Parte II, enero 1994, pp. 175-184.

[ROB94] A. Robinson, "An Application of Recurrent Nets to Phone Probability Estimation", *IEEE Transactions on Neural Networks*, Vol. 5, N° 2, marzo 1994, pp. 298-305.

[ROB98] A. Robinson, "Speech Analysis", ftp://svr-ftp.eng.cam.ac.uk/pub/com.sppech/info, Lent Term 1998.

[ROT93] R. Roth et al., "Large Vocabulary Continuous Speech Recognition of Wall Street Journal Data", *Proc. of ICASSP'93*, Minneapolis, Estados Unidos, 27-30 abril 1993, Vol. II, pp.640-643.

[SAK78] H. Sakoe and S. Chiba, "Dynamic Programming Algorithm Optimization for Spoken Word Recognition", *IEEE Transactions on Acoustic, Speech, and Signal Processing*, Vol. ASSP-26, N° 1, febrero 1978, pp. 43-49.

[SAK92] H. Sakoe, "Dynamic Programming-Based Speech Recognition Algorithms", *Advances in Speech Signal Processing*, S. Furui and M. M. Sondhi editores, Marcel Dekker Inc., 1992, pp. 487-507.

[SAM95] S. Sampath, "Dialectic Specific Features of Australian English Diphthongs in Spontaneus Speech", *Proc. of EUROSPEECH'95, Madrid, septiembre* 1995, pp. 1929-1932.

[SAR98] R. Sarikaya and J. N. Gowdy, "Subband Based Classification of Speech Under Stress", *Proc. of ICASSP'98*, Seattle, Estados Unidos, 12-15 mayo 1998, pp. 569-572.

[SCH84] R. Schwartz et al. "Improved Hidden Markov Modeling of Phonemes for Continuous Speech Recognition", *Proc. of ICASSP'84*, San Diego, Estados Unidos, 19-21 marzo 1984, pp. 35.6.1-35.6.4.

[SCH85] R. Schwartz et al. "Context-Dependent Modeling for Acoustic-Phonetic Recognition of Continuous Speech", *Proc. of ICASSP'85*, pp. 1205-1208.

[SCH87] R. Schwartz, Y. L. Chow and F. Kubala, "Rapid Speaker Adaptation using a Probabilistic Spectral Mapping", *Proc. of ICASSP'87*, Dallas, Estados Unidos, 6-9 abril 1987, pp. 633-636

[SCH89] R. Schulman, "Articulatory dynamics of loud and normal speech", *Journal of Acoustic Society of America*, Vol. 85, N° 1, enero 1989, pp. 295-312.

[SCH90] R. Schwartz and Y. L. Chow, "The N-Best Algorithm: An Efficient and Exact Procedure for Finding the N Most Likely Sentence Hypotheses", *Proc. of ICASSP'90*, Alburquerque, Estados Unidos, 3-6 abril 1990, pp. 81-84.

[SCH92] R. Schwartz et al., "New Uses for the N-Best Sentence Hypotheses within the Byblos Speech Recognition System", *Proc. of ICASSP'92*, Vol. I, pp. 1-4.

[SCH95] P. Schmid and E. Barnard, "Robust, N-best formant tracking", *Proc. of EUROSPEECH'95*, Madrid, septiembre 1995, pp. 737-740.

[SEN88] S. Seneff, "A joint synchrony/mean-rate model of auditory speech processing", *Journal of Phonetics*, Vol. 16, 1988, pp. 55-76.

[SHA88] S. Shamma, "The acoustic features of speech sounds in a model of auditory processing: vowels and voiceless fricatives", *Journal of Phonetics*, Vol. 16, 1988, pp. 77-91.

[SIO95] O. Siohan, "On the Robustness of Linear Discriminant Analysis as a Preprocessing Step for Noisy Speech Recognition", *Proc. of ICASSP'95*, Detroit, Estados Unidos, 9-12 mayo 1995, Detroit, Estados Unidos, 9-12 mayo 1995, pp. 125-128.

[SNE93] R. C. Sneell and F. Milinazzo, "Formant Location from LPC Analysis Data", *IEEE Transactions on Speech and Audio Proc.*, Vol. 1, N° 2, abril 1993, pp. 129-134.

[SOO89] F. K. Soong, "A Phonetically Labeled Acoustic Segment (PLAS) Approach to Speech Analysis-Synthesis", *Proc. of ICASSP'89*, Glasgow, Reino Unido, 23-26 mayo 1989, pp. 584-587.

[SOO91] F. K. Soong and E. F. Huang, "A Tree-Trellis Based Fast Search for Finding the N Best Sentence Hypotheses in Continuous Speech Recognition", *Proc. of ICASSP'91*, Toronto, Canadá, 14-17 mayo 1991, pp. 705-708.

[SUL93] T. M. Sullivan and R. M. Stern, "Multi-Microphone Correlation-Based Processing for Robust Speech Recognition", *Proc. of ICASSP'93*, Minneapolis, Estados Unidos, 27-30 abril 1993, Vol. II, pp. 91-94.

[SUM88] W. Van Summers, "Effects of noise on speech production: Acoustic and perceptual analyses", *Journal of Acoustic Society of America*, Vol. 84, N° 3, septiembre 1988, pp. 917-928.

[TAK92] J. I. Takami and S. Sagayama, "A Successive State Splitting Algorithm for Efficient Allophone Modeling", *Proc. of ICASSP'92*, San Francisco, Estados Unidos, 23-26 marzo 1992, Vol. I, pp. 573-576.

[TEB95] J. Tebelskis, *Speech Recognition using Neural Networks*, Tesis Doctoral, Canegie Mellon University, mayo 1995.

[TEX94] Texas Instruments Incorporated, *TMS320C3x User's Guide*, Digital Signal Processing Products, 1994.

[THO98] D. L. Thomson and R. Chengalvarayan, "Use of Periodicity and Jitter as Speech Recognition Features", *Proc. of ICASSP'98*, Seattle, Estados Unidos, 12-15 mayo 1998, pp. 21-24.

[TIT89] I. R. Titze, "Physiologic and acoustic differences between maleand female voices", *Journal of Acoustic Society of America*, Vol. 85, N° 4, abril 1989, pp. 1699-1707.

[TSE87] H. P. Tseng, M. J. Sabin and E. A. Lee, "Fuzzy Vector Quantization Applied to Hidden Markov Modeling" *Proc. of ICASSP'87*, Dallas, Estados Unidos, 6-9 abril 1987, pp.641-644.

[VAN90] D. Van Compernolle, "Switching Adaptive Filters for Enhancing Noisy and Reverberant Speech from Microphone Array Recordings", *Proc. of ICASSP'90*, Alburquerque, Estados Unidos, 3-6 abril 1990, pp. 833-836.

[VAR90] A. P. Varga and R. K. Moore, "Hidden Markov Model Decomposition of Speech and Noise", *Proc. of ICASSP'90*, Alburquerque, Estados Unidos, 3-6 abril 1990, pp. 845-848.

[VAR97] A. Varnich Hansen, "Acoustic Parameters Optimised for Recognition of Phonetic Features", *Proc. of EUROSPEECH'97*, Rodas, Grecia, septiembre 1997, pp. 397-400.

[VIT67] A. J. Viterbi, "Error Bounds for Convolutional Codes and an Asymptotically Optimum Decoding Algorithm", *IEEE Transactions on Information Theory*, Vol. IT-13, N°. 2, abril 1967, pp. 260-309.

[WAI88] A. Waibel et al., "Phoneme Recognition: Neural Networks vs. Hidden Markov Models", *Proc. of ICASSP'88*, Nueva York, Estados Unidos, 11-14 abril 1988, pp. 107-110.

[WAI89] A. Waibel, T. Hanazawa, G. E. Hinton, K. Shikano and K. J. Lang, "Phoneme recognition usign time-delay neural networks", *IEEE Transactions on Acoustic, Speech, and Signal Processing*, Vol. ASSP-37, N° 3, marzo 1989, pp. 328-339.

[WAK77] H. Wakita, "Normalization of Vowels by Vocal-Tract Length and Its Application to Vowel Identification", *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. ASSP-25, N°. 2, abril 1977, pp. 183-192.

[WAN73] M. D. Wang and R. C. Bilger, "Consonant confusions in noise: a study of perceptual features", *Journal of Acoustic Society of America*, Vol. 54, N° 5, junio 1973, pp. 1248-1266.

[WEL87] C. J. Wellekens, "Explicit Correlation in Hidden Markov Models for Speech Recognition", *Proc. of ICASSP'87*, Dallas, Estados Unidos, 6-9 abril 1987, pp. 384-386.

[WEL95] L. Welling and H. Ney, "A Model for Efficient Formant Estimation", *Proc. of ICASSP'95*, Detroit, Estados Unidos, 9-12 mayo 1995, pp. 797-800.

[WEL98a] L. Welling and H. Ney, "Formant Estimation for Speech Recognition", *IEEE Transactions on Speech and Audio Processing*, Vol. 6, N° 1, enero 1998, pp. 36-48.

[WEL98b] L. Welling et al., "A Study on Speaker Normalization Using Vocal Tract Normalization and Speaker Adaptive Training", *Proc. of ICASSP'98*, Seattle, Estados Unidos, 12-15 mayo 1998, Seattle, Estados Unidos, 12-15 mayo 1998, pp. 797-800.

[WIL72] C. E. Williams and K. N. Stevens, "Emotions and Speech: Some Acoustical Correlates", *Journal of Acoustic Society of America*, Vol. 52, N° 4, marzo 1972, pp. 1238-1250.

[WIL96] J. G. Wilpon and C. N. Jacobsen, "A Study of Speech Recognition for Children and the Elderly", *Proc. of ICASSP'96*, Atlanta, Estados Unidos, 7-10 mayo 1996, pp. 349-352.

[WOO94] P. C. Woodland et al., "Large Vocabulary Continuous Speech Recognition Using HTK", *Proc. of ICASSP'94*, Adelaida, Australia, 19-22 abril 1994, Vol. II, pp. 125-128.

[YAN95] R. Yang, M. Majaniemi and P. Haavisto, "Dynamic Parameter Compensation for Speech Recognition in Noise", *Proc. of EUROSPEECH'95, Madrid, septiembre* 1995, pp. 469-472.

[YOU96] S. Young, "A Review of Large-vocabulary Continuous-speech Recognition", *IEEE Signal Processing Magazine*, septiembre 1996, pp. 45-57.

[ZAV94] G. Zavaliagkos et al., "A Hybrid Segmental Neural Net/Hidden Markov Model System for Continuous Speech Recognition", *IEEE Transactions on Speech and Audio Processing*, Vol. 2, N°. 1, Parte II, enero 1994, pp. 151-160.

[ZHA97] Y. Zhao, "Overcoming Speaker Variability in Automatic Speech Recognition. The Speaker Adaptation Approach", *Talker Variability in Speech Processing*, K. Johnson and J. W. Mullenix editores, Academic Press Limited, 1997, pp. 191-209.

[ZWI80] E. Zwicker and E. Terhardt, "Analytical expressions for critical-band rate and critical bandwidth as a function of frequency", *Journal of Acoustic Society of America*, Vol. 68, N° 5, febrero 1980, pp. 1523-1525.

APÉNDICE A. RESUMEN DE RESULTADOS DE LOS EXPERIMENTOS CON DIFERENTES PARAMETRIZADORES

A.1. EXPERIMENTOS BASADOS EN LPC

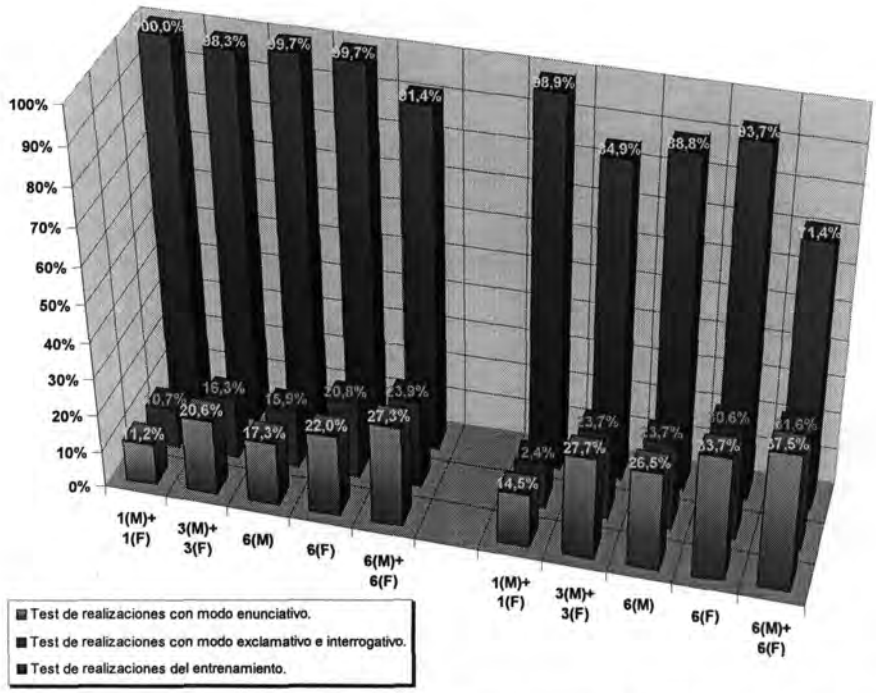

Figura A.1. Tasas de reconocimiento para el experimento LPC-1a (LPC-Coeficientes *a* [16 coeficientes]). La parte de la izquierda corresponde a los entrenamientos efectuados solamente con realizaciones con modo enunciativo.

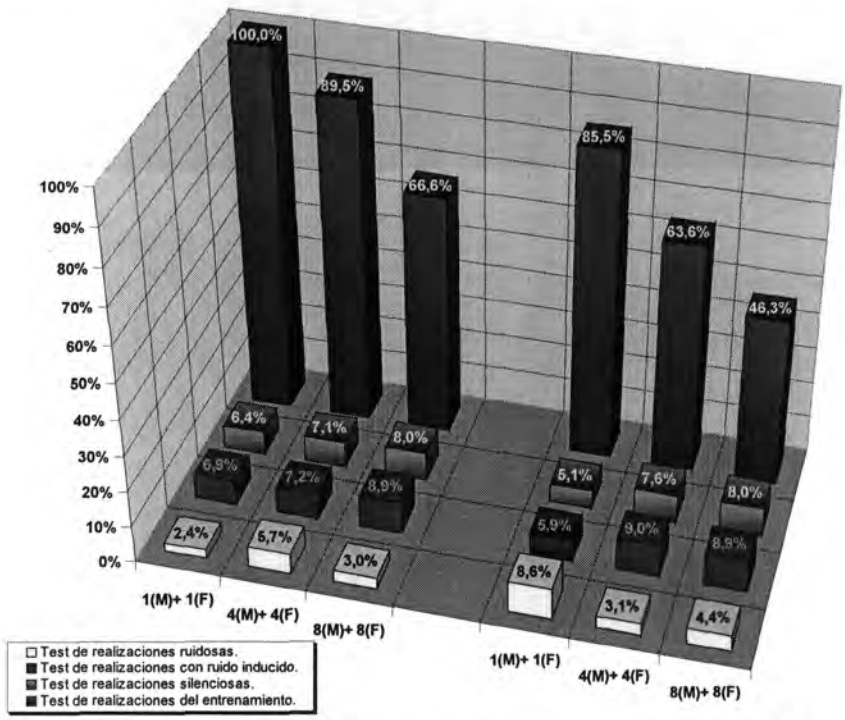

Figura A.2. Tasas de reconocimiento para el experimento LPC-1b (LPC-Coeficientes *a* [16 coeficientes]). La parte de la izquierda corresponde a los entrenamientos efectuados solamente con realizaciones silenciosas.

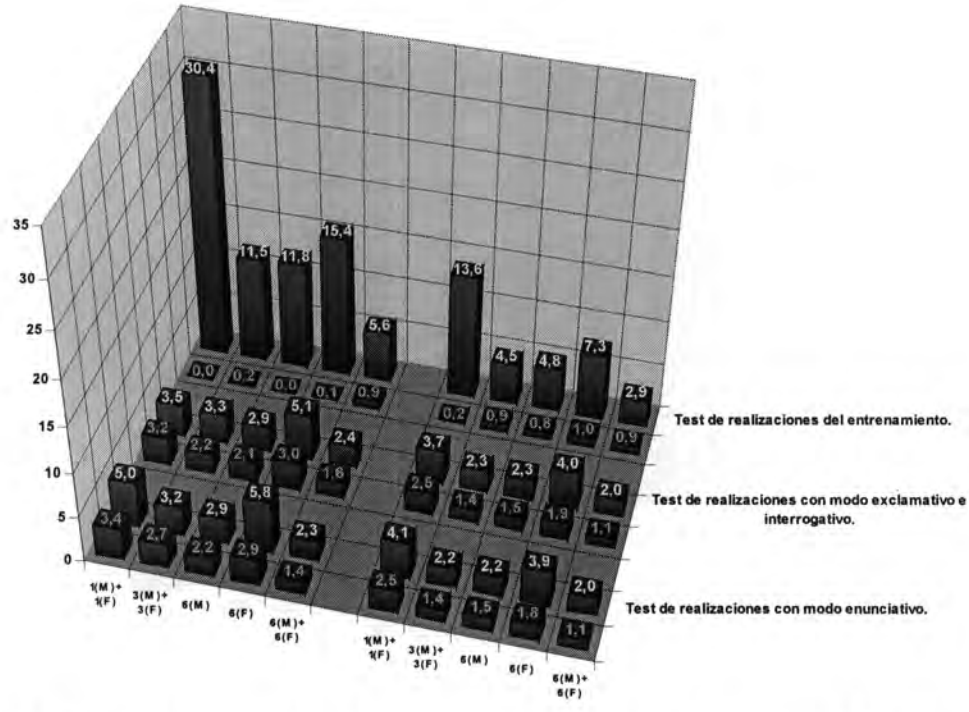

Figura A.3. Tasas de fiabilidad para el experimento LPC-1a (LPC-Coeficientes *a* [16 coeficientes]). En verde aparecen las tasas para los casos reconocidos correctamente. En rojo los casos clasificados erróneamente.

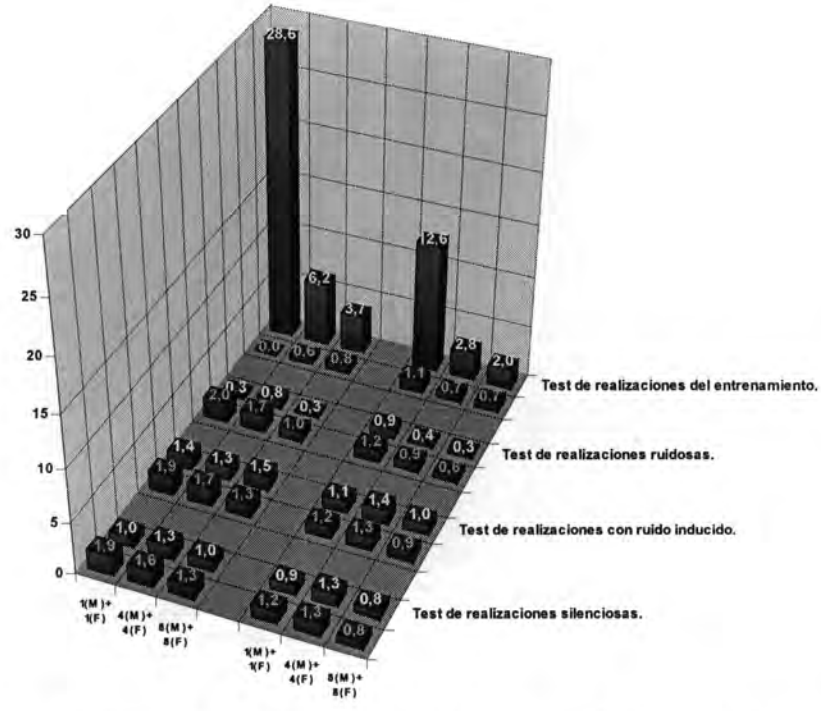

Figura A.4. Tasas de fiabilidad para el experimento LPC-1b (LPC-Coeficientes *a* [16 coeficientes]). En verde aparecen las tasas para los casos reconocidos correctamente. En rojo los casos clasificados erróneamente.

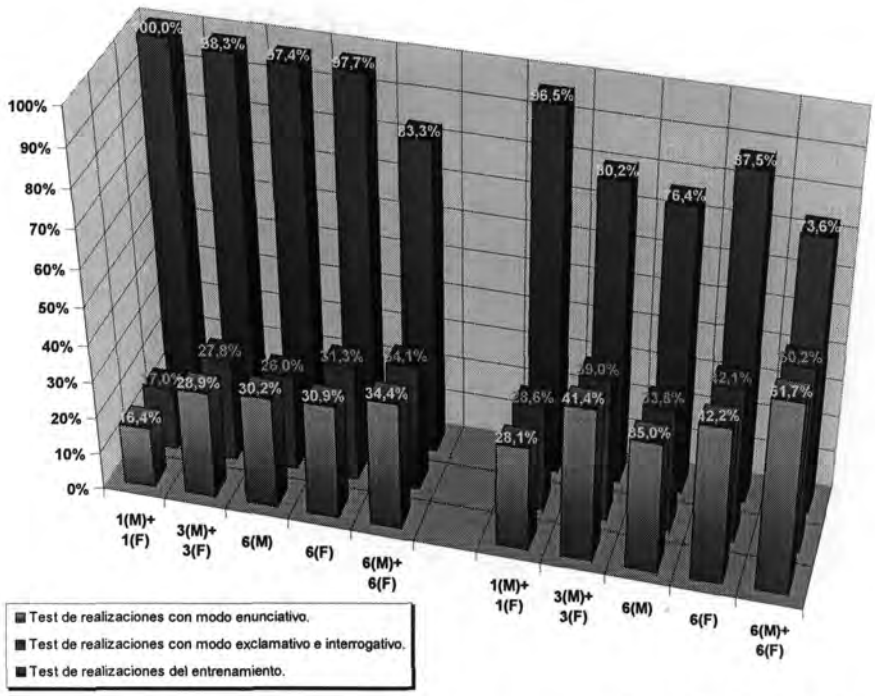

Figura A.5. Tasas de reconocimiento para el experimento LPC-2a (LPC-Coeficientes *a* [16 coeficientes + 16 coeficientes anteriores]). La parte de la izquierda corresponde a los entrenamientos efectuados solamente con realizaciones con modo enunciativo.

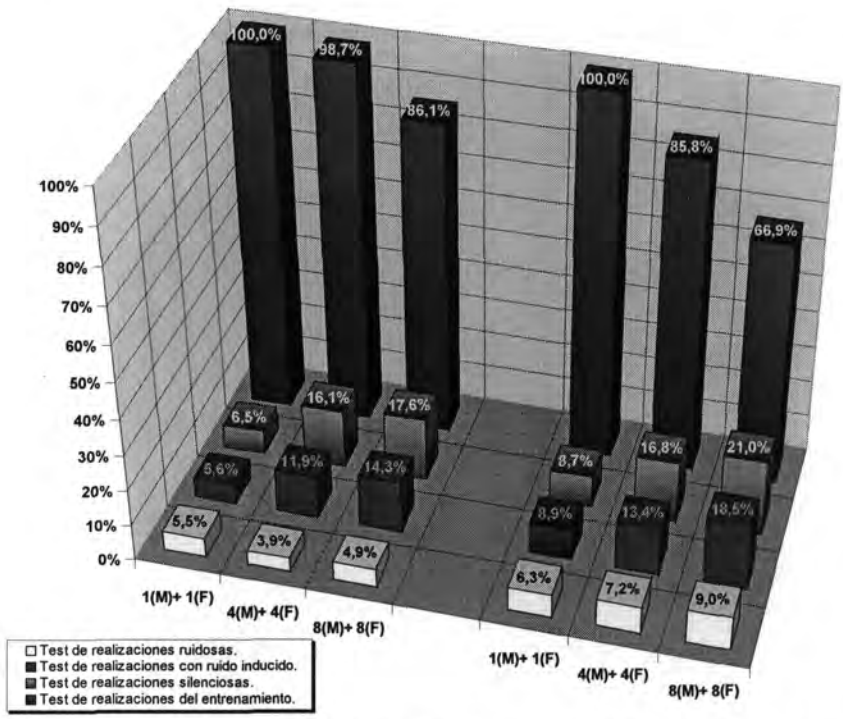

Figura A.6. Tasas de reconocimiento para el experimento LPC-2b (LPC-Coeficientes *a* [16 coeficientes + 16 coeficientes anteriores]). La parte de la izquierda corresponde a los entrenamientos efectuados solamente con realizaciones silenciosas.

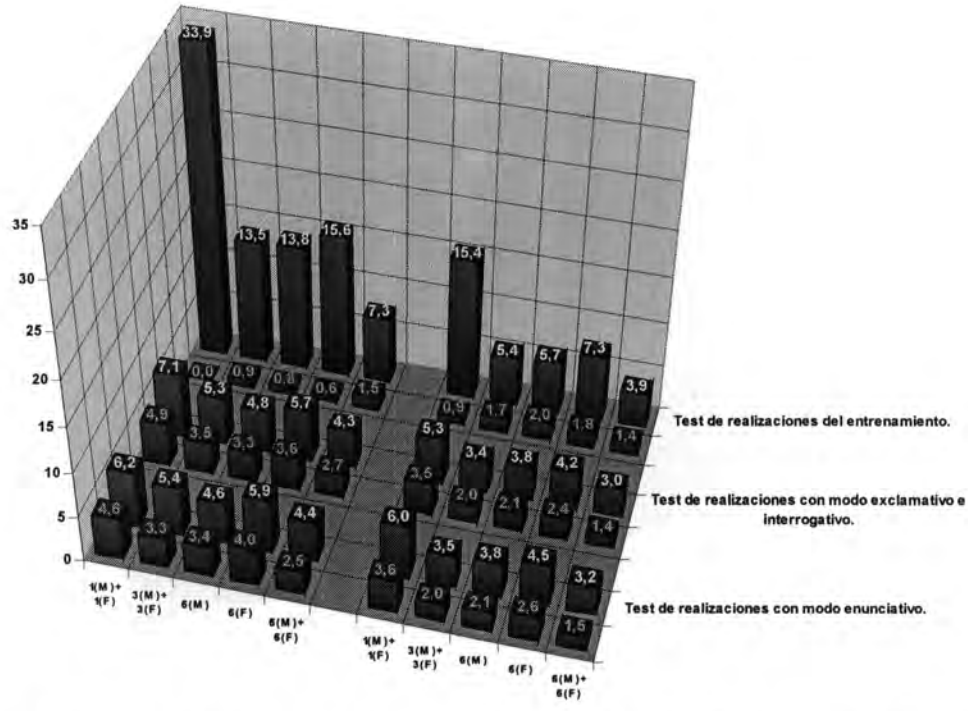

Figura A.7. Tasas de fiabilidad para el experimento LPC-2a (LPC-Coeficientes *a* [16 coeficientes + 16 coeficientes anteriores]). En verde aparecen las tasas para los casos reconocidos correctamente. En rojo los casos clasificados erróneamente.

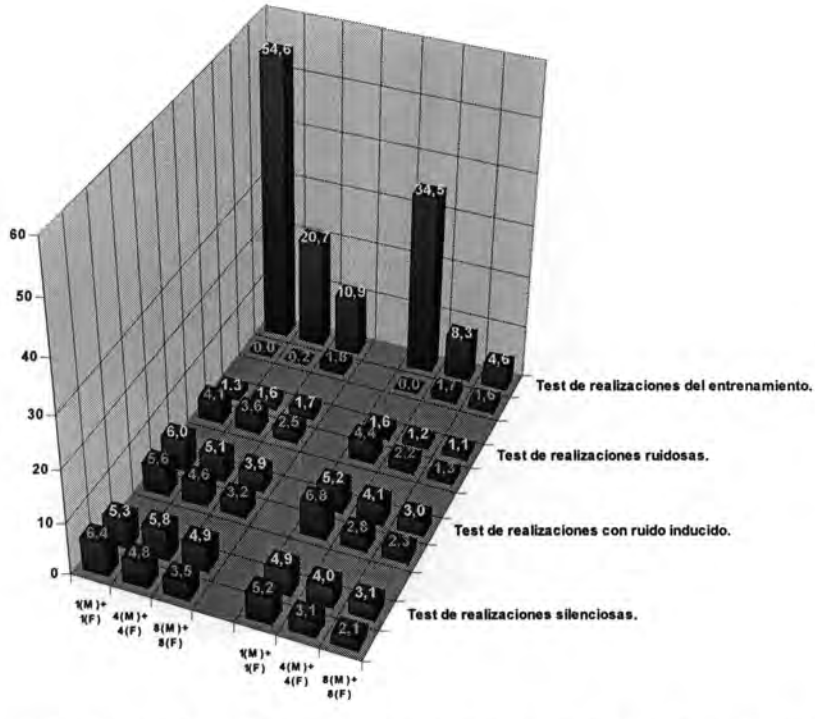

Figura A.8. Tasas de fiabilidad para el experimento LPC-2b (LPC-Coeficientes *a* [16 coeficientes + 16 coeficientes anteriores]). En verde aparecen las tasas para los casos reconocidos correctamente. En rojo los casos clasificados erróneamente.

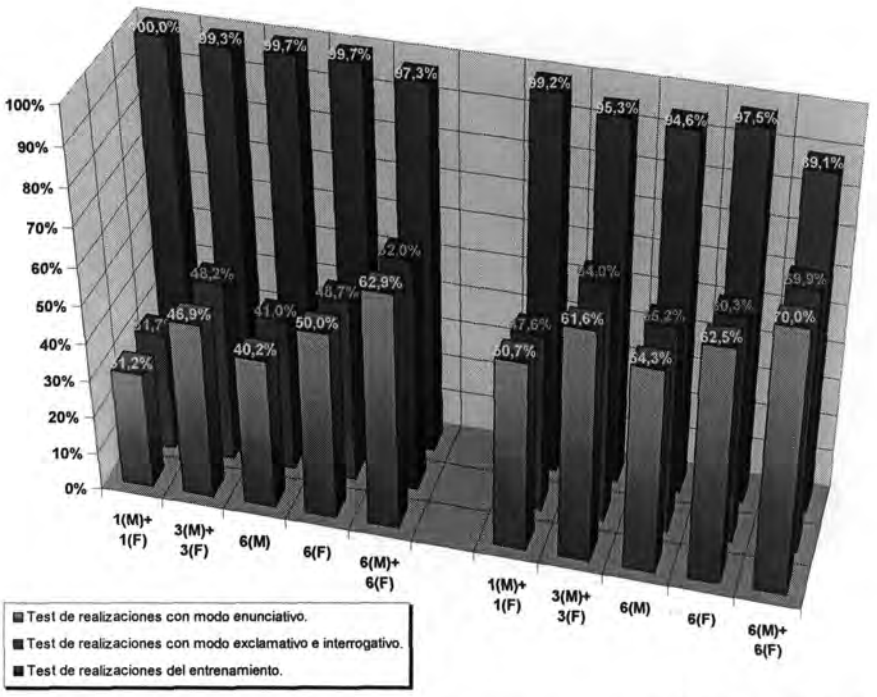

Figura A.9. Tasas de reconocimiento para el experimento LPC-3a (LPC-Cepstrum [10 coeficientes]). La parte de la izquierda corresponde a los entrenamientos efectuados solamente con realizaciones con modo enunciativo.

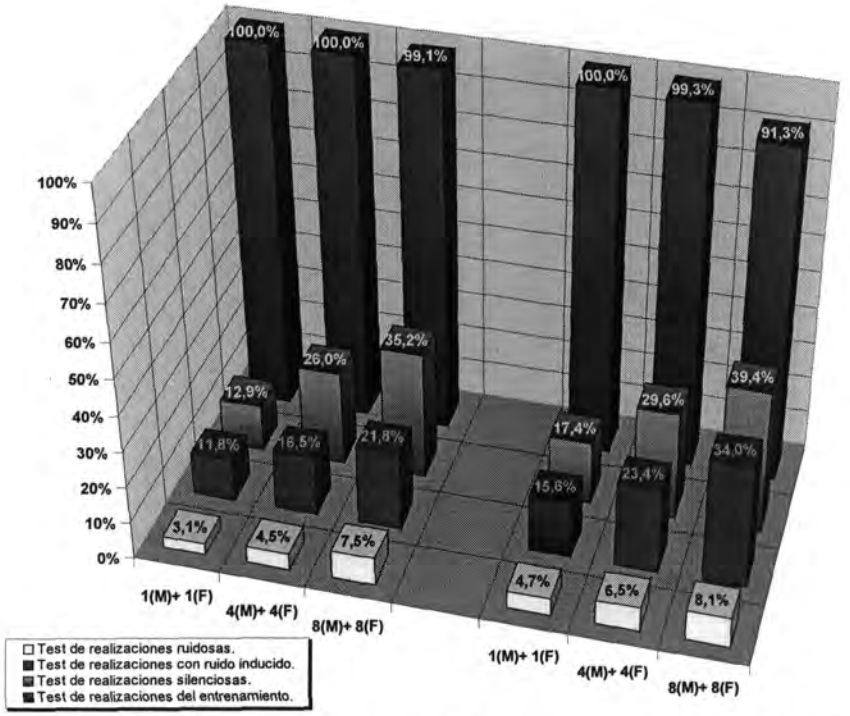

Figura A.10. Tasas de reconocimiento para el experimento LPC-3b (LPC-Cepstrum [10 coeficientes]). La parte de la izquierda corresponde a los entrenamientos efectuados solamente con realizaciones silenciosas.

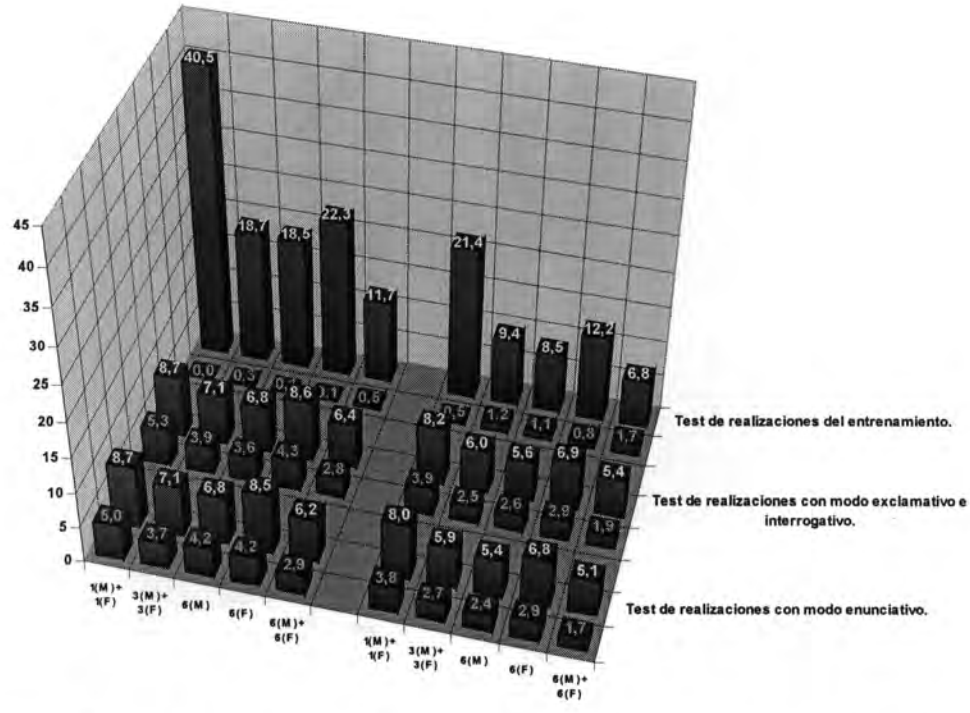

Figura A.11. Tasas de fiabilidad para el experimento LPC-3a (LPC-Cepstrum [10 coeficientes]). En verde aparecen las tasas para los casos reconocidos correctamente. En rojo los casos clasificados erróneamente.

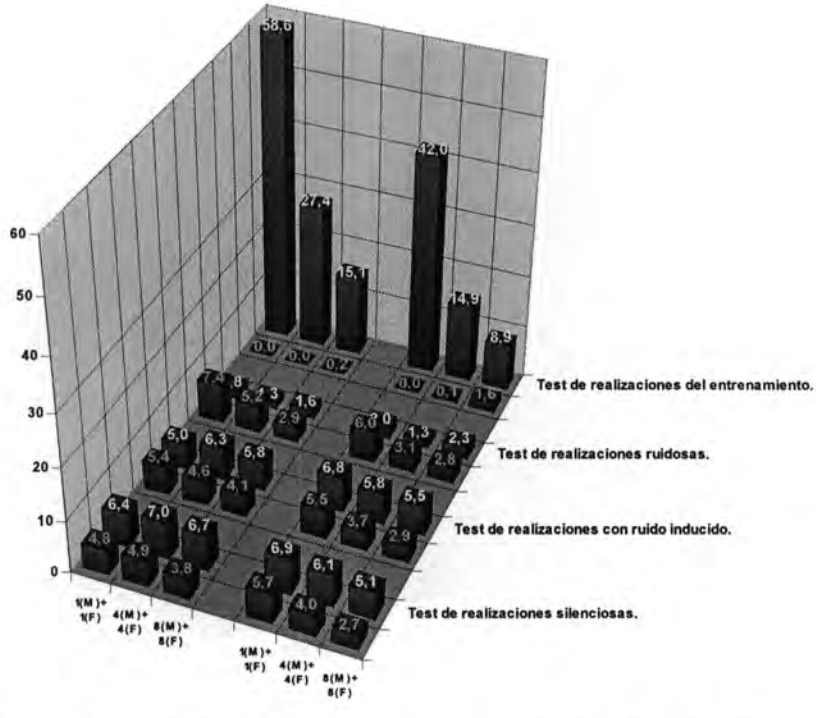

Figura A.12. Tasas de fiabilidad para el experimento LPC-3b (LPC-Cepstrum [10 coeficientes]). En verde aparecen las tasas para los casos reconocidos correctamente. En rojo los casos clasificados erróneamente.

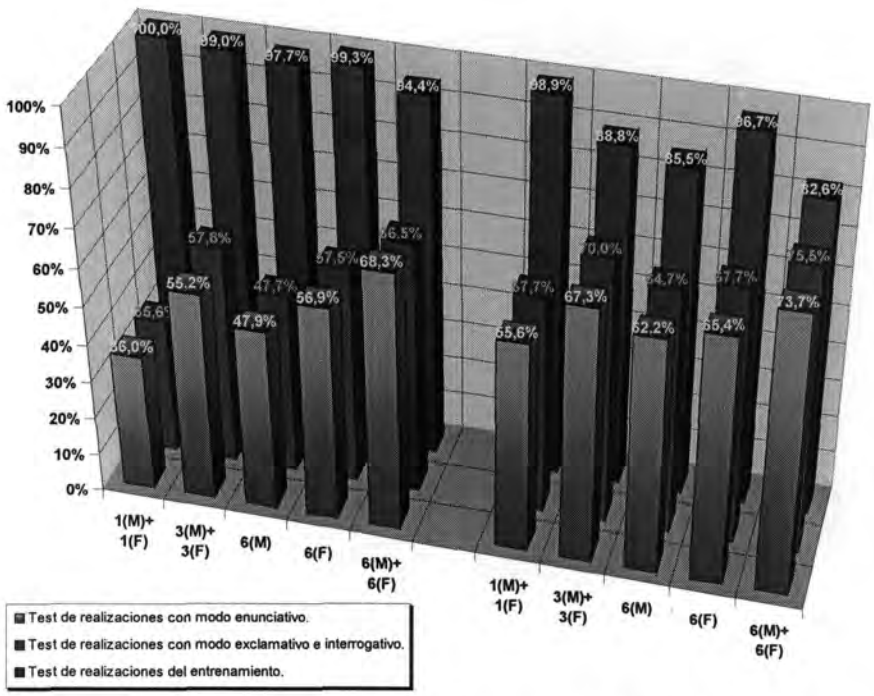

Figura A.13. Tasas de reconocimiento para el experimento LPC-4a (LPC-Cepstrum [10 coeficientes + 10 coeficientes delta]). La parte de la izquierda corresponde a los entrenamientos efectuados solamente con realizaciones con modo enunciativo.

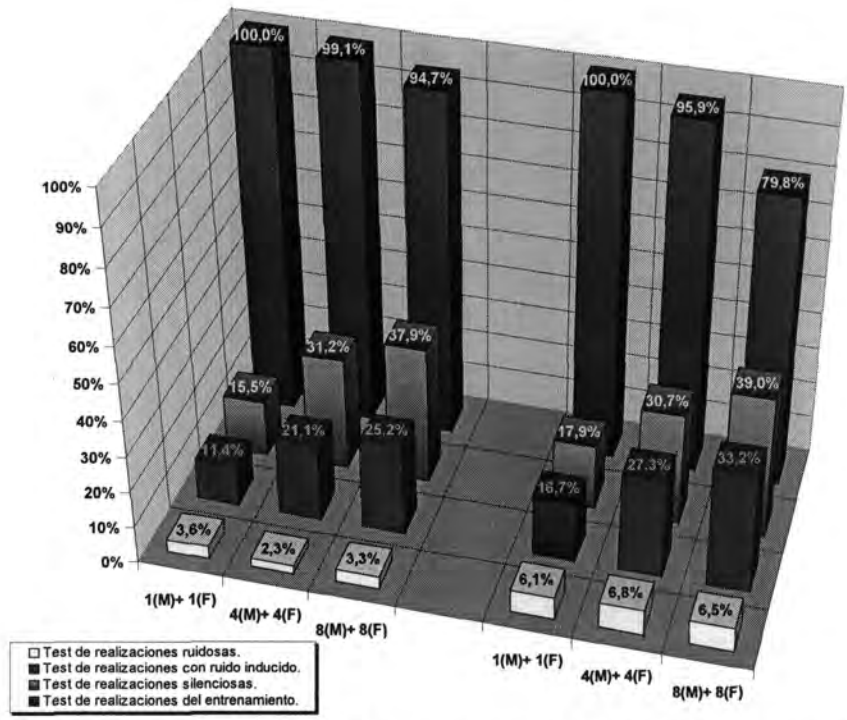

Figura A.14. Tasas de reconocimiento para el experimento LPC-4b (LPC-Cepstrum [10 coeficientes + 10 coeficientes delta]). La parte de la izquierda corresponde a los entrenamientos efectuados solamente con realizaciones silenciosas.

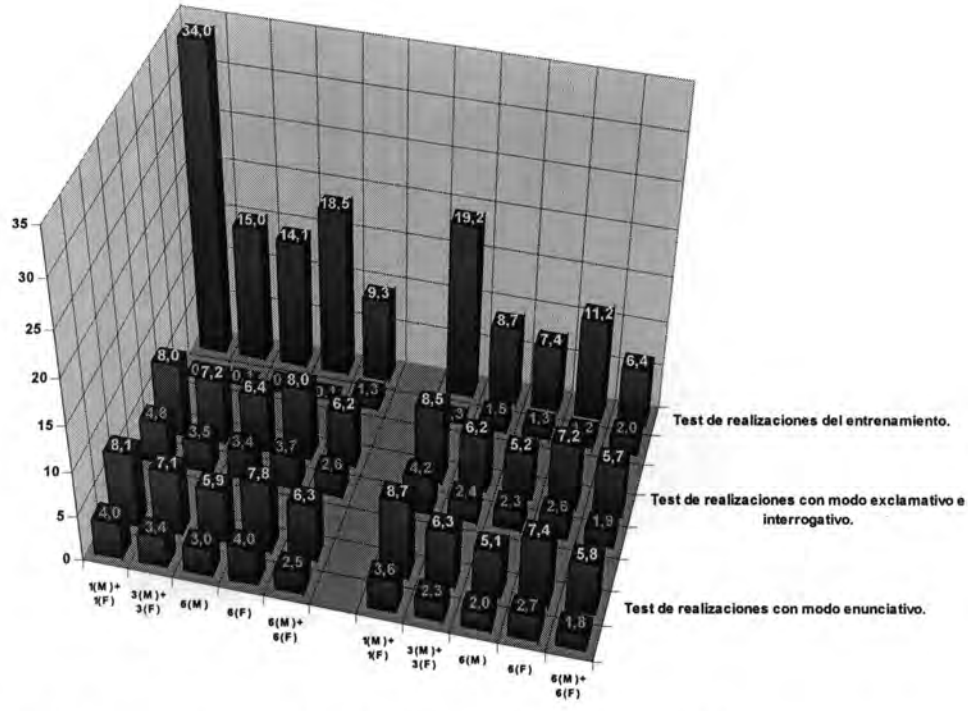

Figura A.15. Tasas de fiabilidad para el experimento LPC-4a (LPC-Cepstrum [10 coeficientes + 10 coeficientes delta]). En verde aparecen las tasas para los casos reconocidos correctamente. En rojo los casos clasificados erróneamente.

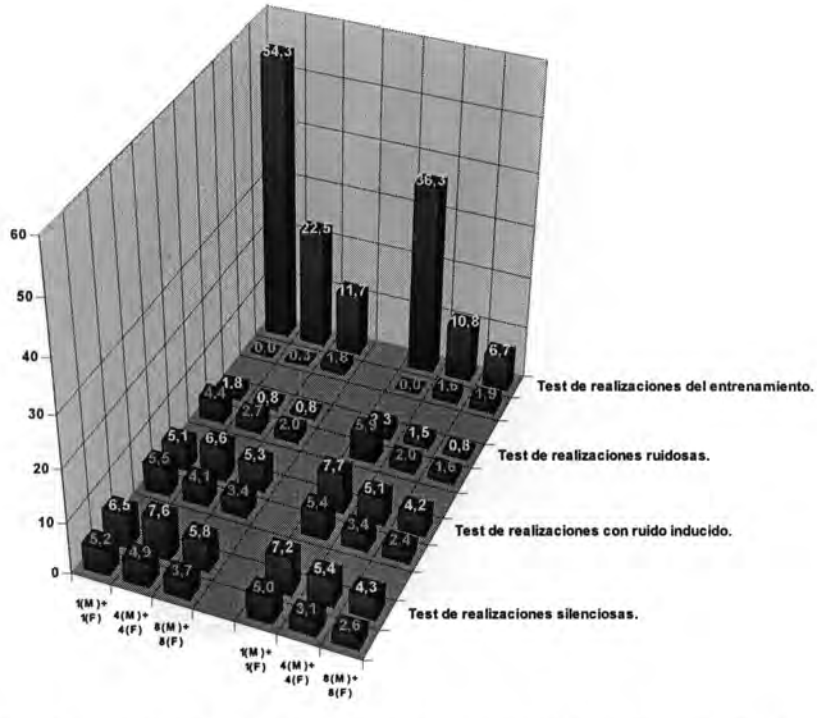

Figura A.16. Tasas de fiabilidad para el experimento LPC-4b (LPC-Cepstrum [10 coeficientes + 10 coeficientes delta]). En verde aparecen las tasas para los casos reconocidos correctamente. En rojo los casos clasificados erróneamente.

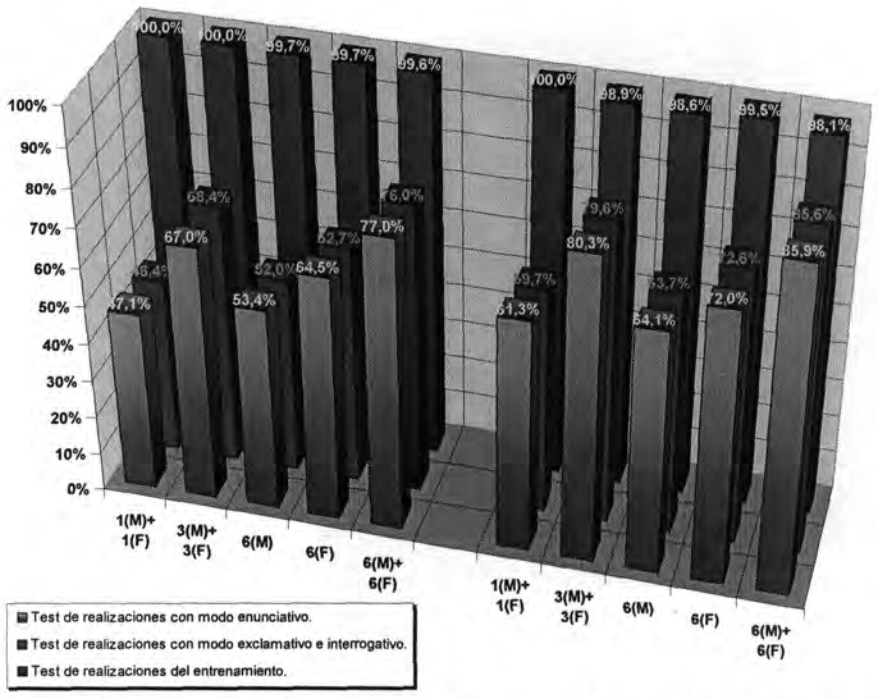

Figura A.17. Tasas de reconocimiento para el experimento LPC-5a (LPC-Bandas Energía [20 bandas]). La parte de la izquierda corresponde a los entrenamientos efectuados solamente con realizaciones con modo enunciativo.

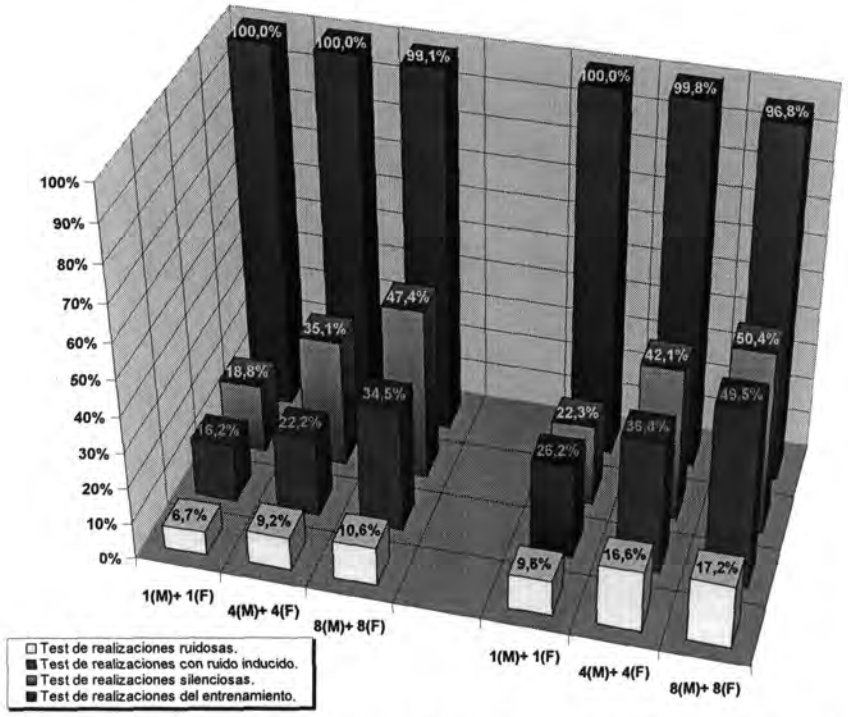

Figura A.18. Tasas de reconocimiento para el experimento LPC-5b (LPC-Bandas Energía [20 bandas]). La parte de la izquierda corresponde a los entrenamientos efectuados solamente con realizaciones silenciosas.

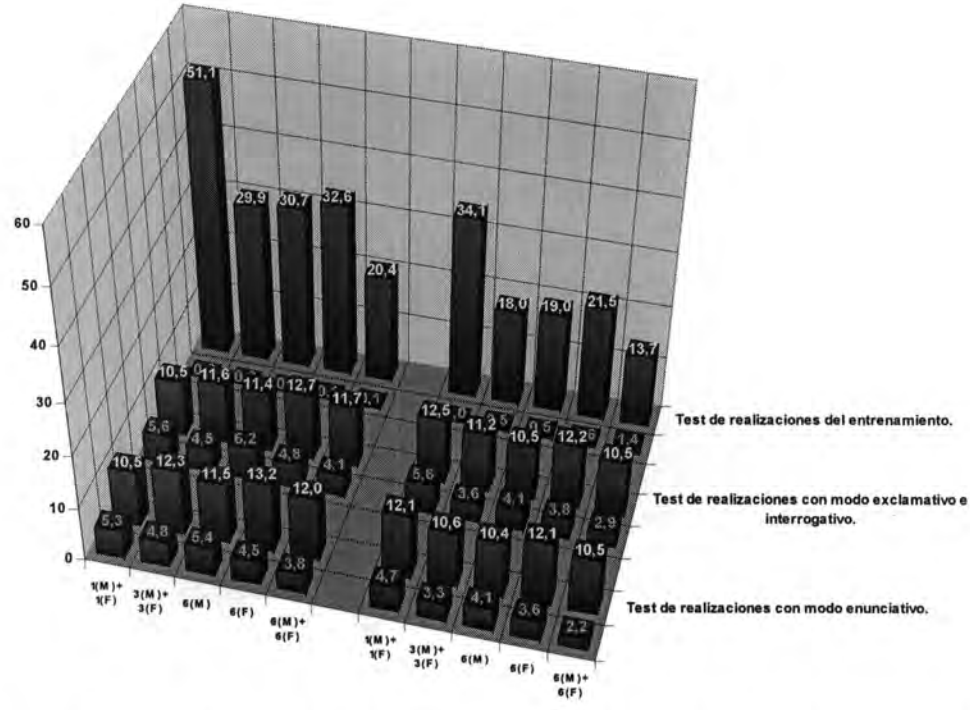

Figura A.19. Tasas de fiabilidad para el experimento LPC-5a (LPC-Bandas Energía [20 bandas]). En verde aparecen las tasas para los casos reconocidos correctamente. En rojo los casos clasificados erróneamente.

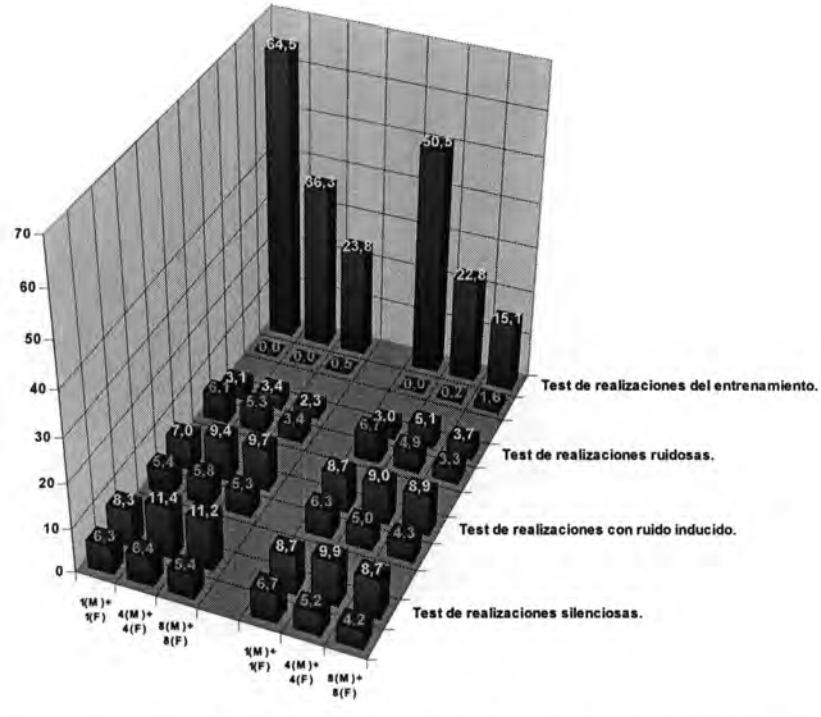

Figura A.20. Tasas de fiabilidad para el experimento LPC-5b (LPC-Bandas Energía [20 bandas]). En verde aparecen las tasas para los casos reconocidos correctamente. En rojo los casos clasificados erróneamente.

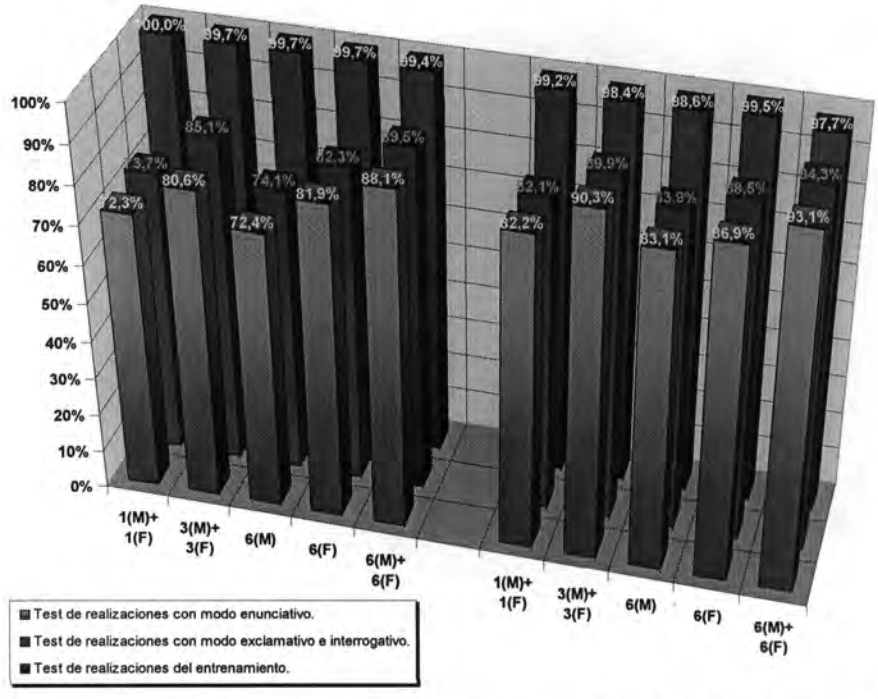

Figura A.21. Tasas de reconocimiento para el experimento LPC-6a (LPC-Bandas Energía [20 bandas + 20 bandas delta <5 tramas>]). La parte de la izquierda corresponde a los entrenamientos efectuados solamente con realizaciones con modo enunciativo.

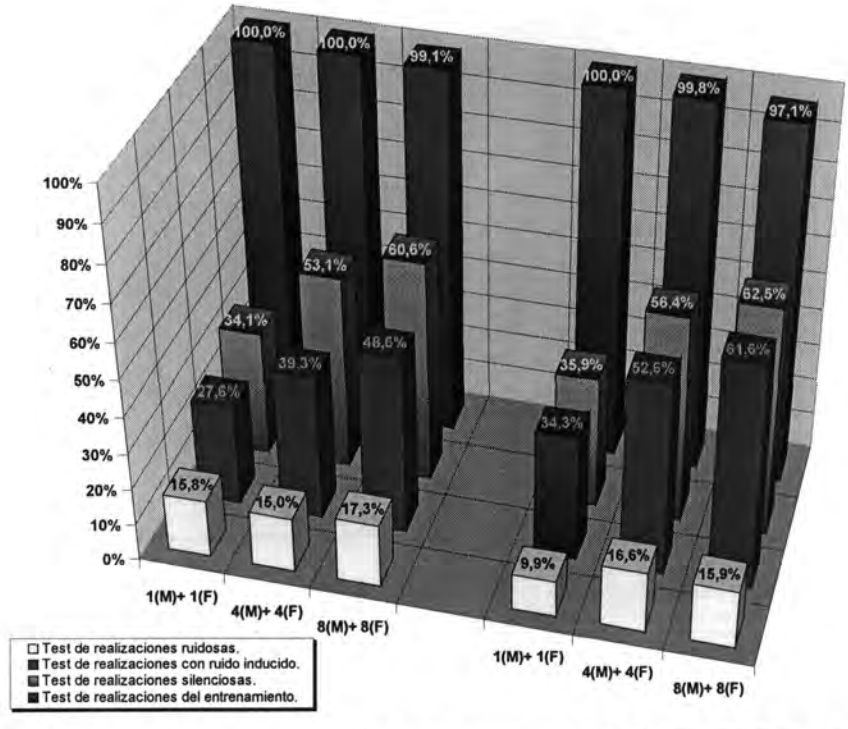

Figura A.22. Tasas de reconocimiento para el experimento LPC-6b (LPC-Bandas Energía [20 bandas + 20 bandas delta <5 tramas>]). La parte de la izquierda corresponde a los entrenamientos efectuados solamente con realizaciones silenciosas.

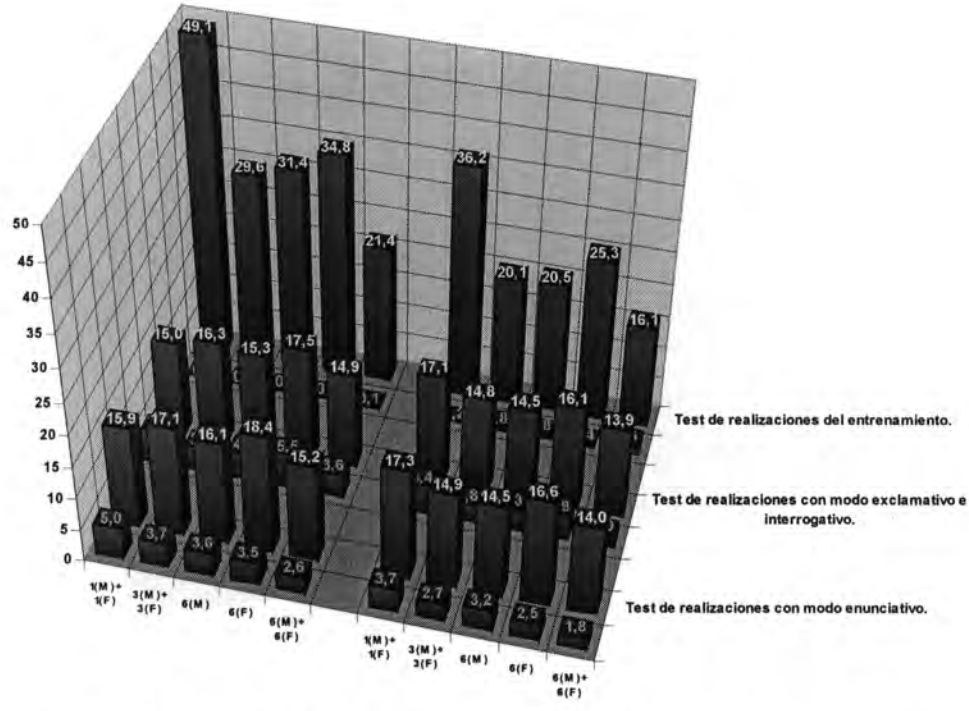

Figura A.23. Tasas de fiabilidad para el experimento LPC-6a (LPC-Bandas Energía [20 bandas + 20 bandas delta <5 tramas>]). En verde aparecen las tasas para los casos reconocidos correctamente. En rojo los casos clasificados erróneamente.

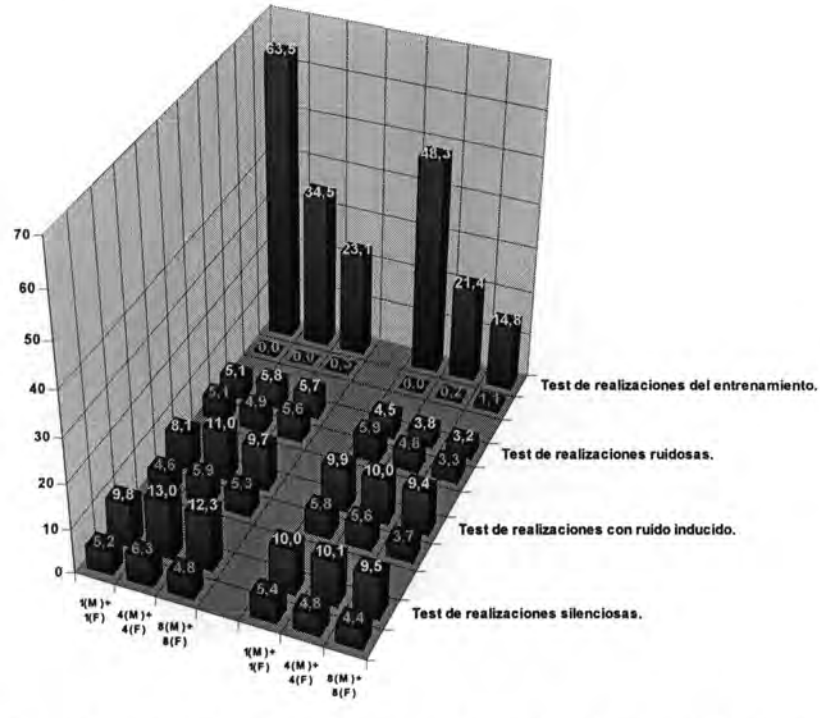

Figura A.24. Tasas de fiabilidad para el experimento LPC-6b (LPC-Bandas Energía [20 bandas + 20 bandas delta <5 tramas>]). En verde aparecen las tasas para los casos reconocidos correctamente. En rojo los casos clasificados erróneamente.

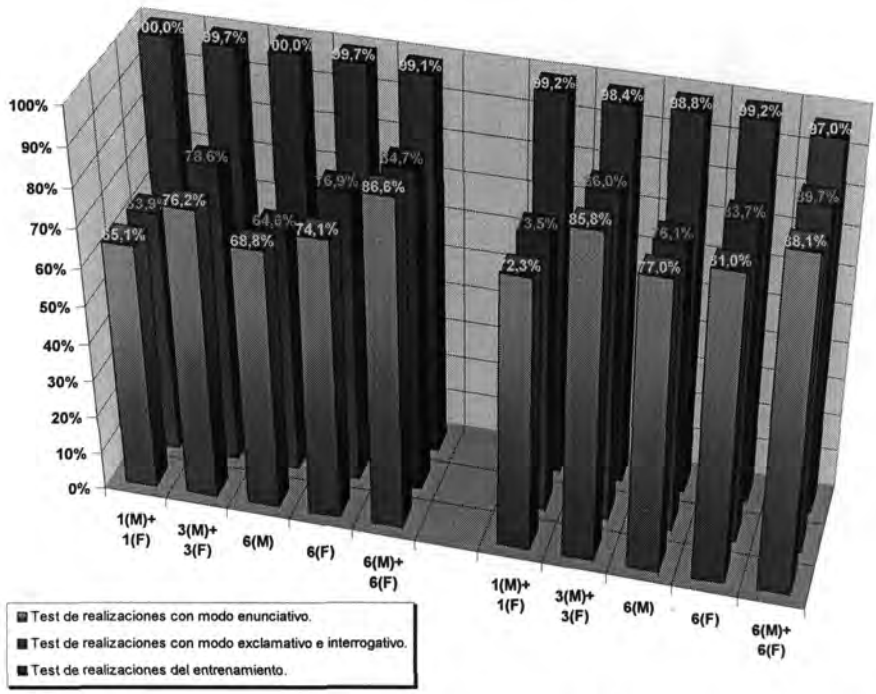

Figura A.25. Tasas de reconocimiento para el experimento LPC-7a (LPC-Bandas Energía [20 bandas + 20 bandas diferencia <2 tramas>]). La parte de la izquierda corresponde a los entrenamientos efectuados solamente con realizaciones con modo enunciativo.

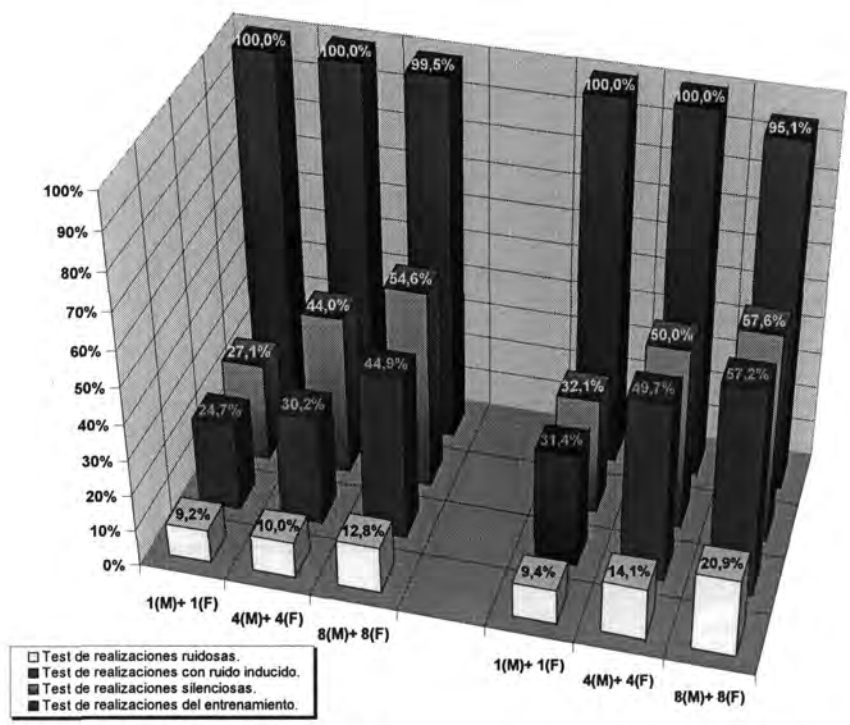

Figura A.26. Tasas de reconocimiento para el experimento LPC-7b (LPC-Bandas Energía [20 bandas + 20 bandas diferencia <2 tramas>]). La parte de la izquierda corresponde a los entrenamientos efectuados solamente con realizaciones silenciosas.

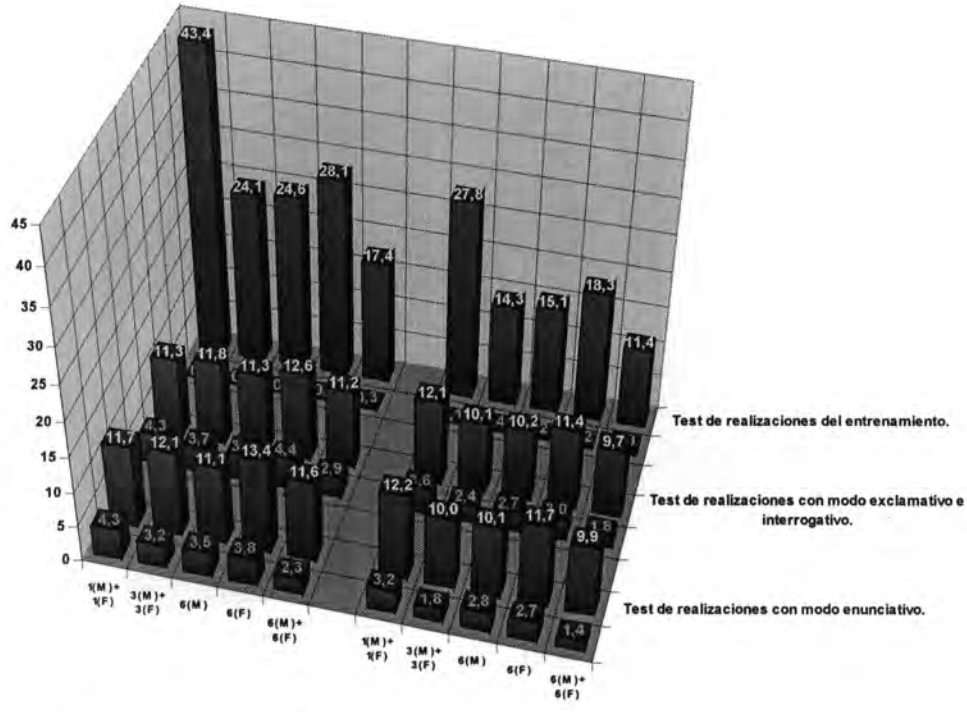

Figura A.27. Tasas de fiabilidad para el experimento LPC-7a (LPC-Bandas Energía [20 bandas + 20 bandas diferencia <2 tramas>]). En verde aparecen las tasas para los casos reconocidos correctamente. En rojo los casos clasificados erróneamente.

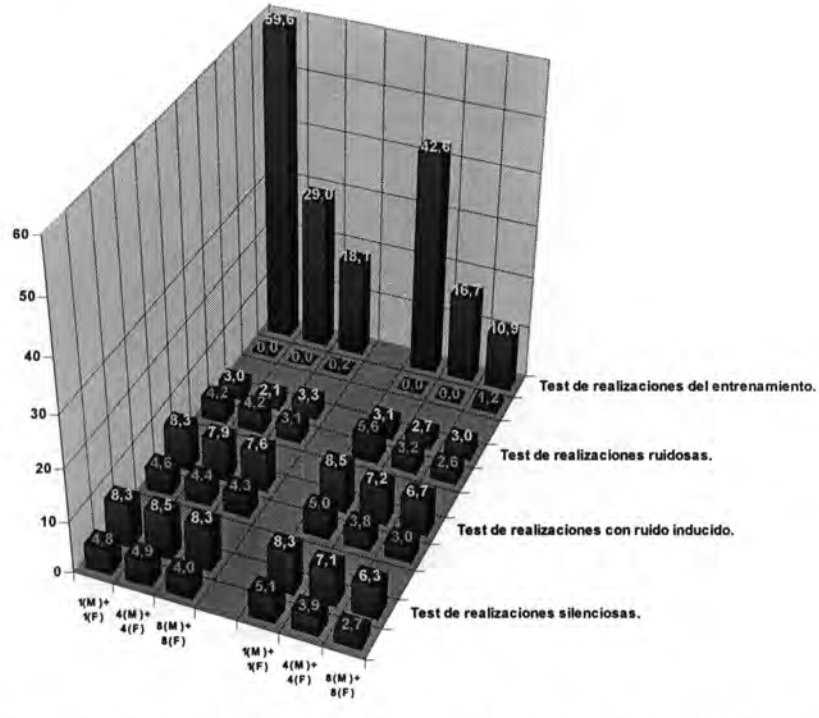

Figura A.28. Tasas de fiabilidad para el experimento LPC-7b (LPC-Bandas Energía [20 bandas + 20 bandas diferencia <2 tramas>]). En verde aparecen las tasas para los casos reconocidos correctamente. En rojo los casos clasificados erróneamente.

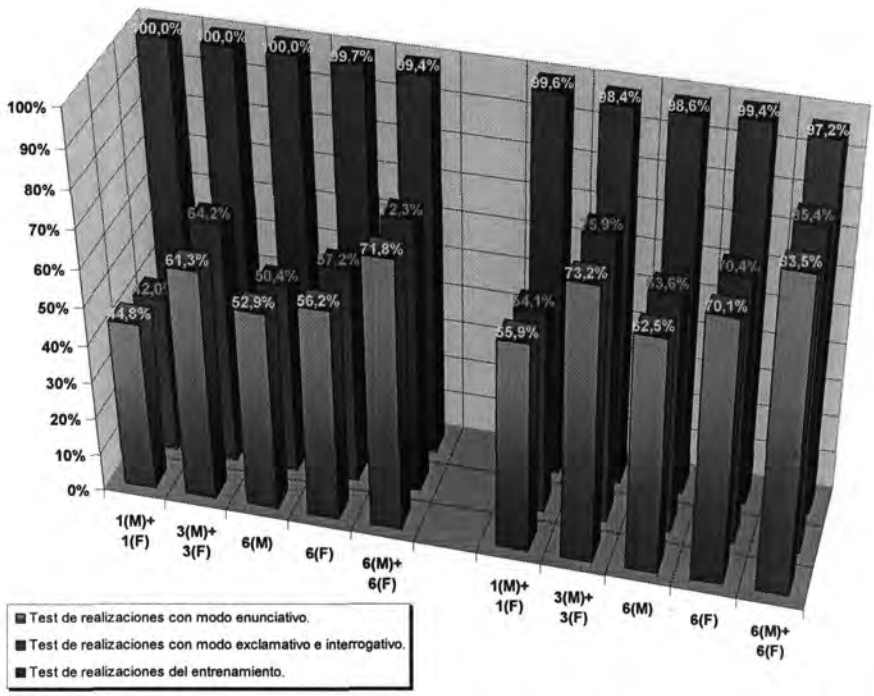

Figura A.29. Tasas de reconocimiento para el experimento LPC-8a. La parte de la izquierda corresponde a los entrenamientos efectuados solamente con realizaciones con modo enunciativo.

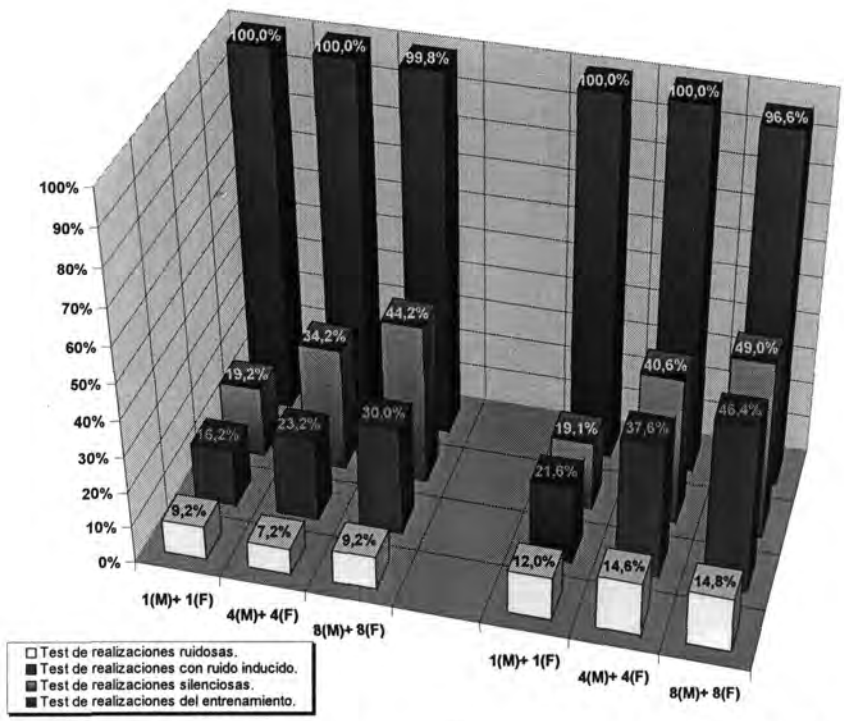

Figura A.30. Tasas de reconocimiento para el experimento LPC-8b (LPC-Bandas Energía Normalizadas [20 bandas]). La parte de la izquierda corresponde a los entrenamientos efectuados solamente con realizaciones silenciosas.

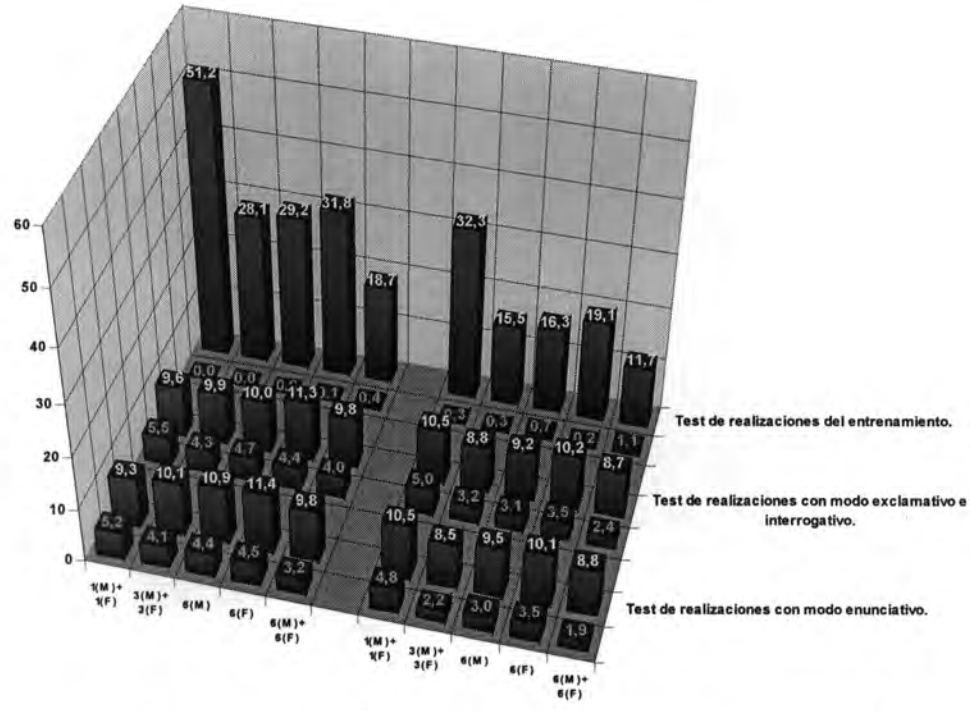

Figura A.31. Tasas de fiabilidad para el experimento LPC-8a. En verde aparecen las tasas para los casos reconocidos correctamente. En rojo los casos clasificados erróneamente.

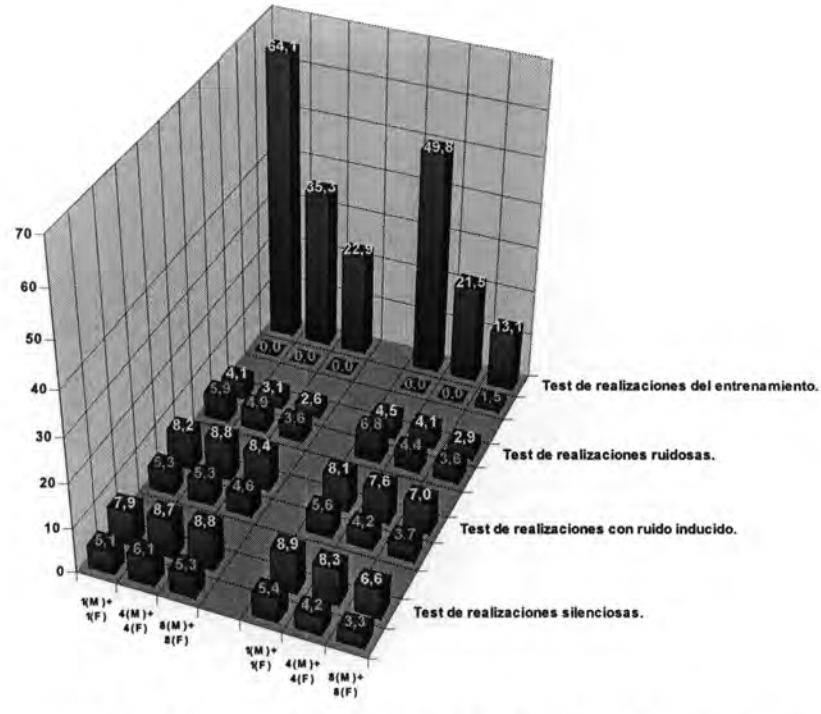

Figura A.32. Tasas de fiabilidad para el experimento LPC-8b (LPC-Bandas Energía Normalizadas [20 bandas]). En verde aparecen las tasas para los casos reconocidos correctamente. En rojo los casos clasificados erróneamente.

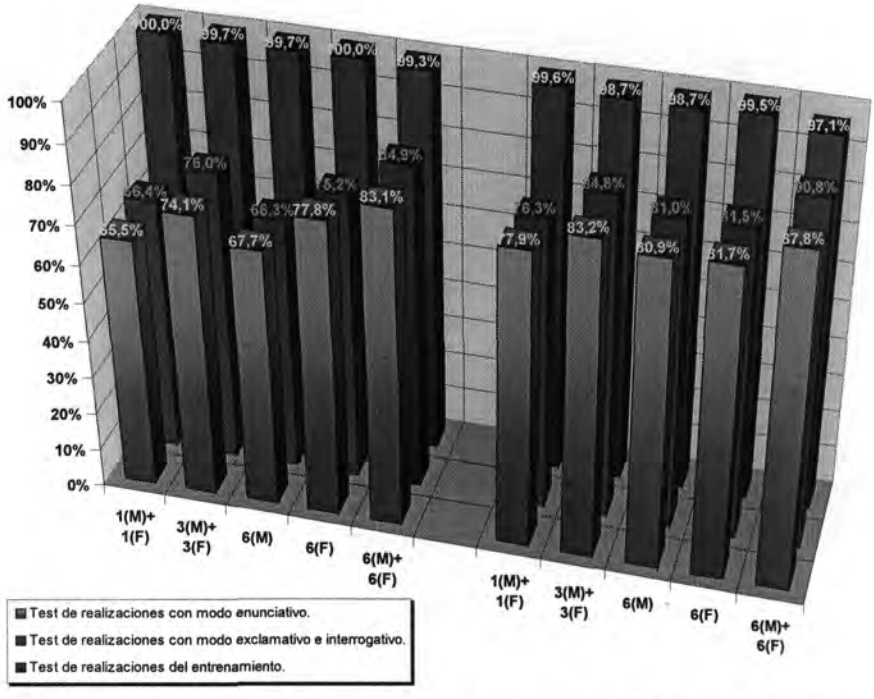

Figura A.33. Tasas de reconocimiento para el experimento LPC-9a (LPC-Bandas Energía Normalizadas [20 bandas + 20 bandas delta <5 tramas>]). La parte de la izquierda corresponde a los entrenamientos efectuados solamente con realizaciones con modo enunciativo.

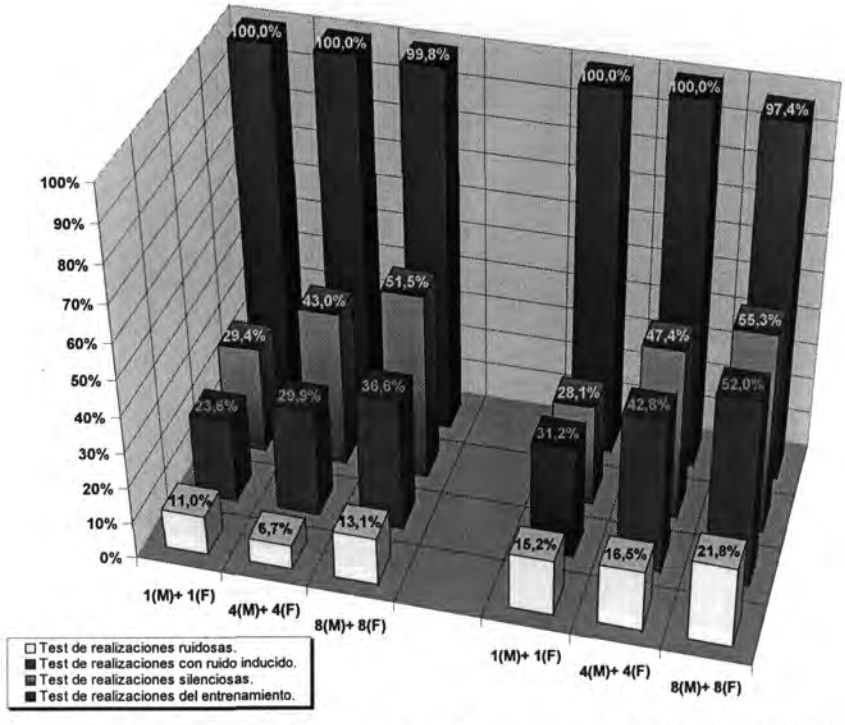

Figura A.34. Tasas de reconocimiento para el experimento LPC-9b (LPC-Bandas Energía Normalizadas [20 bandas + 20 bandas delta <5 tramas>]). La parte de la izquierda corresponde a los entrenamientos efectuados solamente con realizaciones silenciosas.

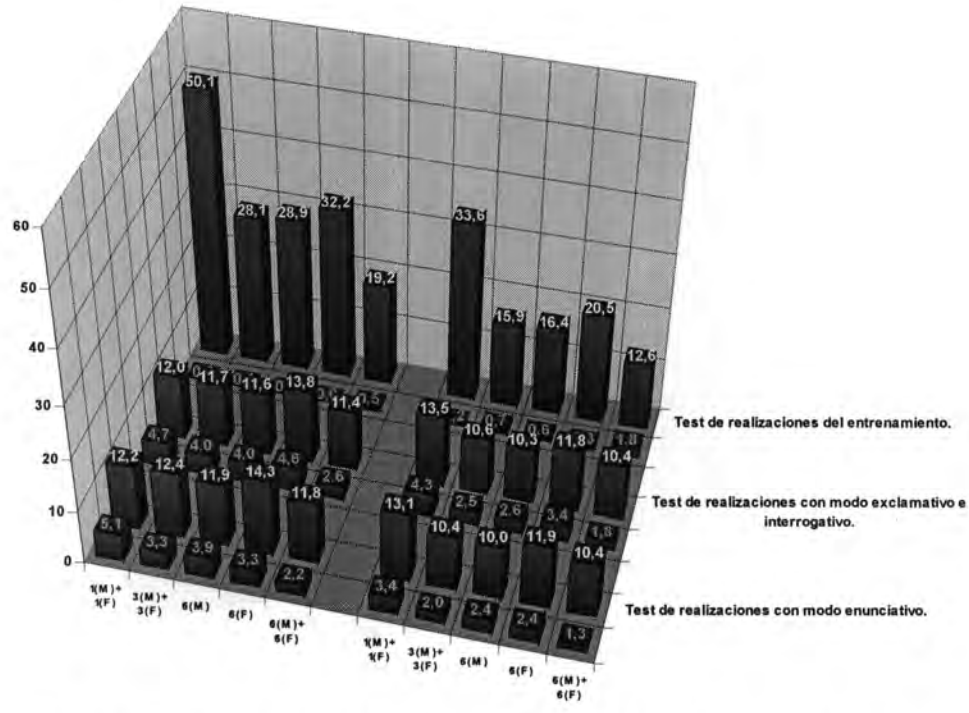

Figura A.35. Tasas de fiabilidad para el experimento LPC-9a (LPC-Bandas Energía Normalizadas [20 bandas + 20 bandas delta <5 tramas>]). En verde aparecen las tasas para los casos reconocidos correctamente. En rojo los casos clasificados erróneamente.

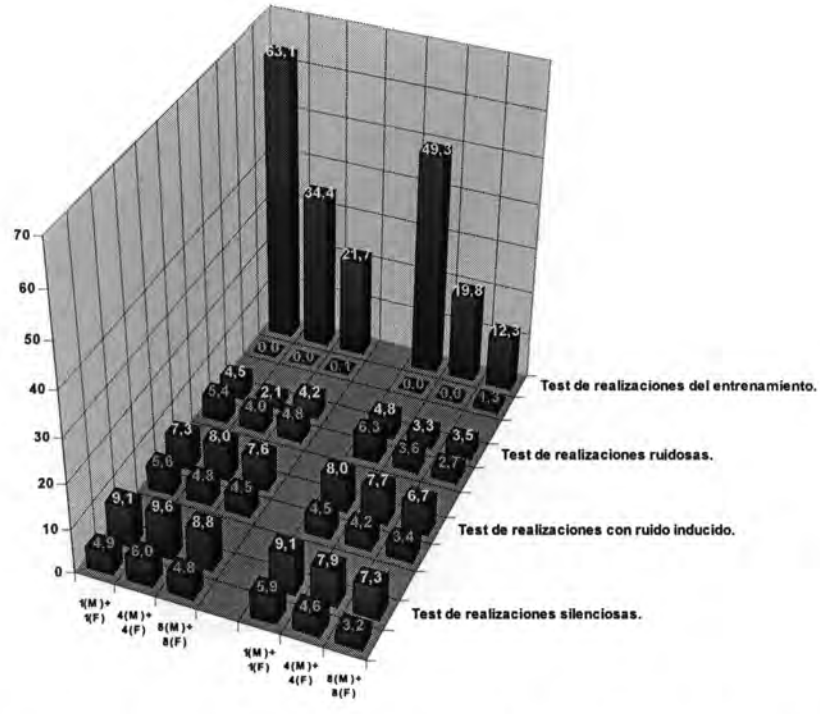

Figura A.36. Tasas de fiabilidad para el experimento LPC-9b (LPC-Bandas Energía Normalizadas [20 bandas + 20 bandas delta <5 tramas>]). En verde aparecen las tasas para los casos reconocidos correctamente. En rojo los casos clasificados erróneamente.

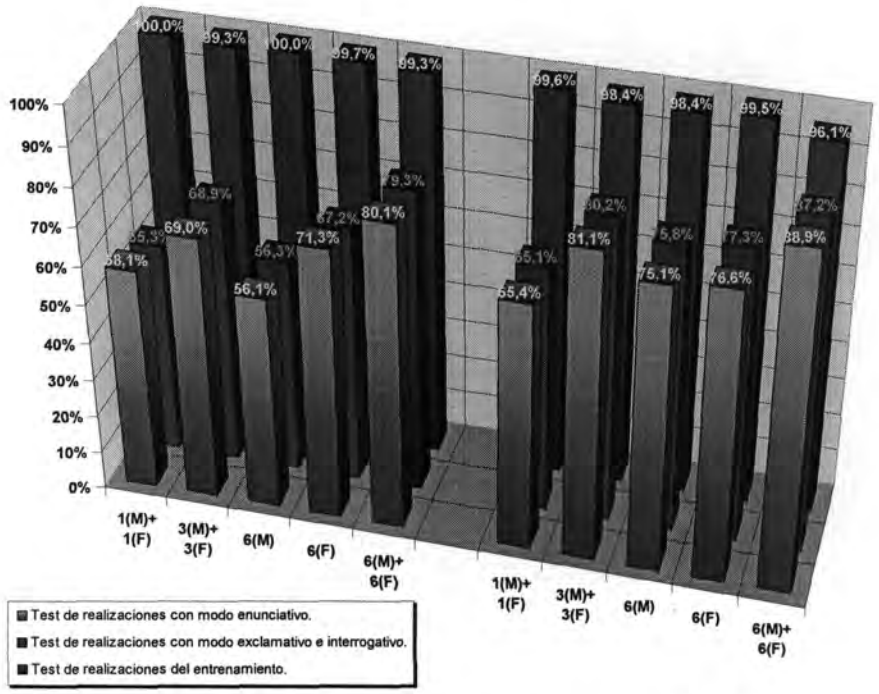

Figura A.37. Tasas de reconocimiento para el experimento LPC-10a (LPC-Bandas Energía Normalizadas [20 bandas + 20 bandas diferencia <2 tramas>]). La parte de la izquierda corresponde a los entrenamientos efectuados solamente con realizaciones con modo enunciativo.

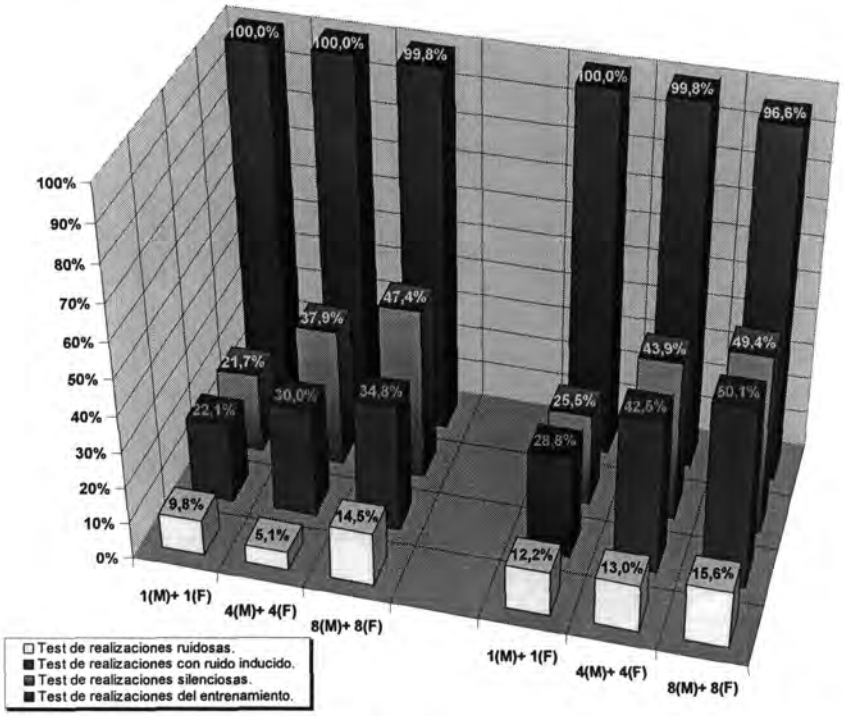

Figura A.38. Tasas de reconocimiento para el experimento LPC-10b (LPC-Bandas Energía Normalizadas [20 bandas + 20 bandas diferencia <2 tramas>]). La parte de la izquierda corresponde a los entrenamientos efectuados solamente con realizaciones silenciosas.

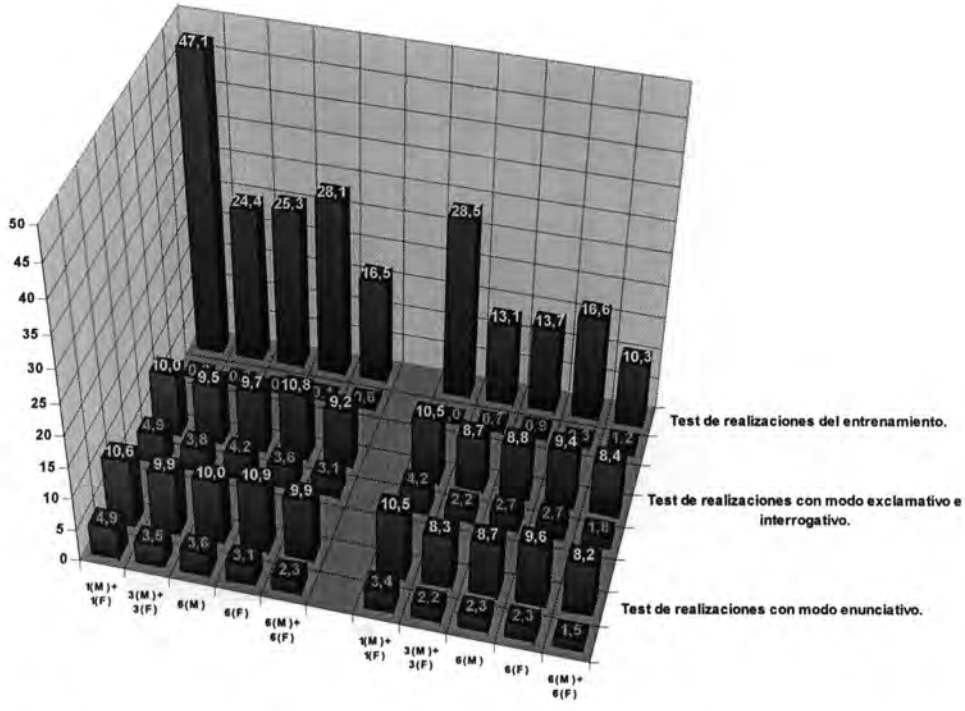

Figura A.39. Tasas de fiabilidad para el experimento LPC-10a (LPC-Bandas Energía Normalizadas [20 bandas + 20 bandas diferencia <2 tramas>]). En verde aparecen las tasas para los casos reconocidos correctamente. En rojo los casos clasificados erróneamente.

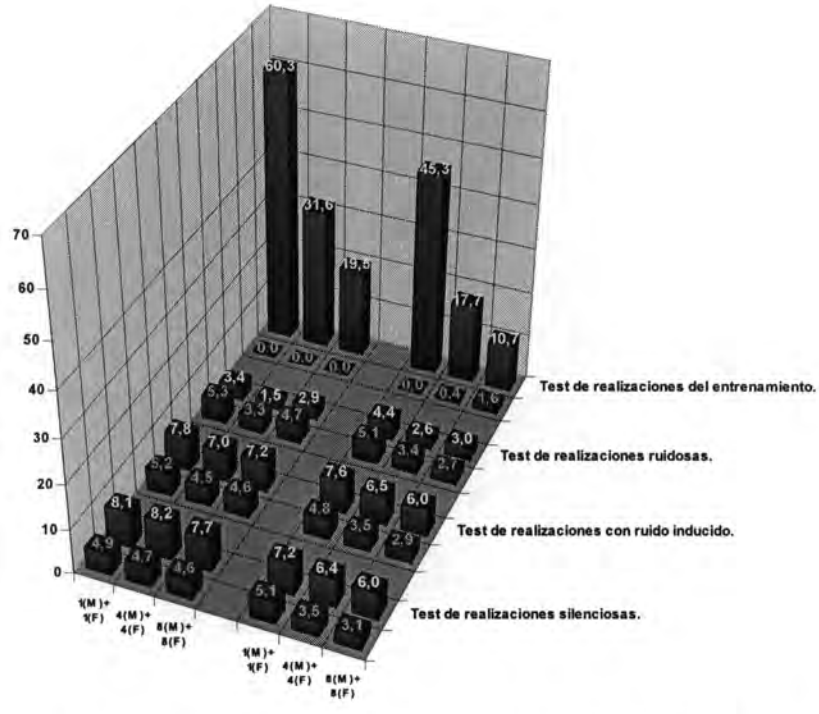

Figura A.40. Tasas de fiabilidad para el experimento LPC-10b (LPC-Bandas Energía Normalizadas [20 bandas + 20 bandas diferencia <2 tramas>]). En verde aparecen las tasas para los casos reconocidos correctamente. En rojo los casos clasificados erróneamente.

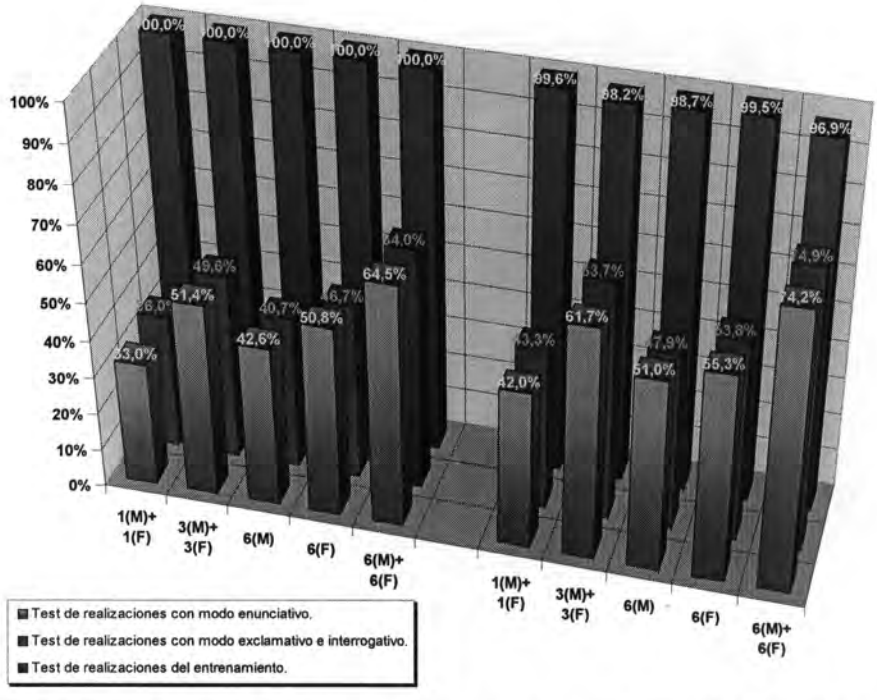

Figura A.41. Tasas de reconocimiento para el experimento LPC-11a (LPC-Bandas Energía Cepstrum [10 coeficientes]). La parte de la izquierda corresponde a los entrenamientos efectuados solamente con realizaciones con modo enunciativo.

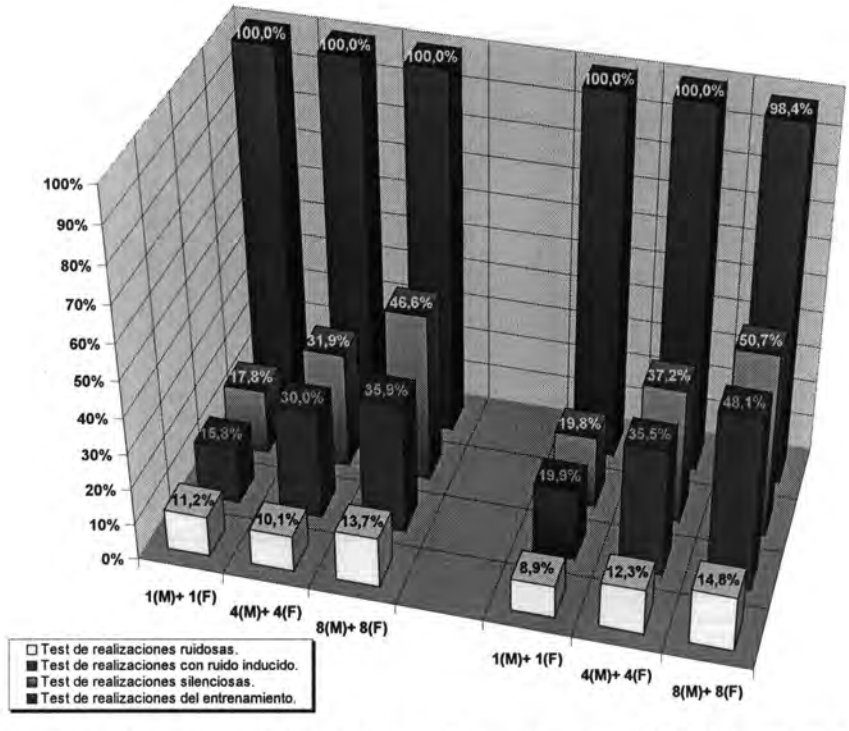

Figura A.42. Tasas de reconocimiento para el experimento LPC-11b (LPC-Bandas Energía Cepstrum [10 coeficientes]). La parte de la izquierda corresponde a los entrenamientos efectuados solamente con realizaciones silenciosas.

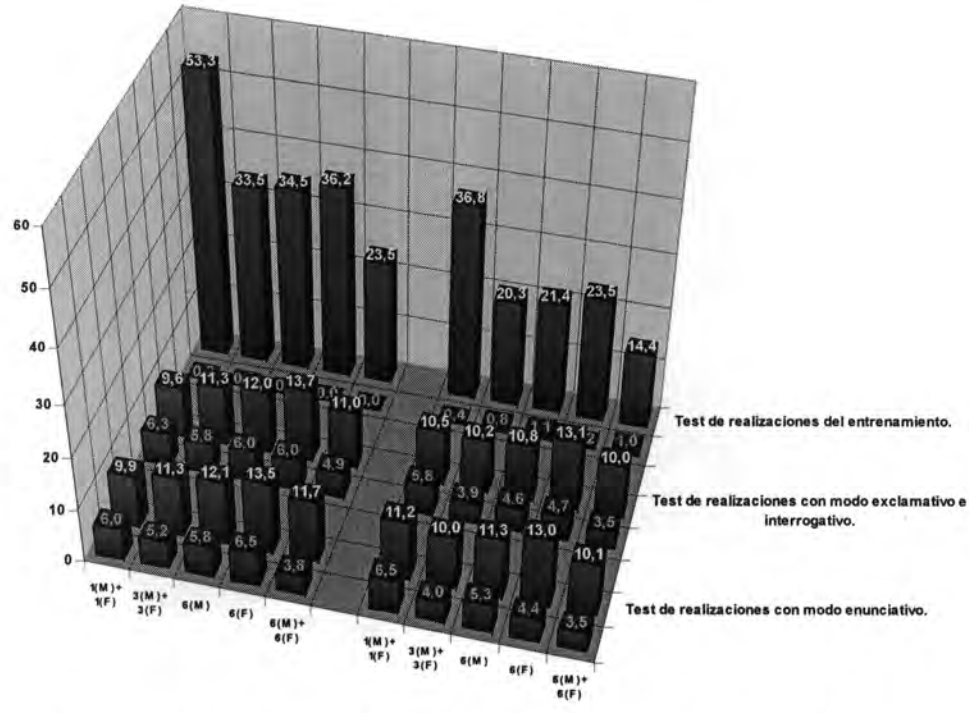

Figura A.43. Tasas de fiabilidad para el experimento LPC-11a (LPC-Bandas Energía Cepstrum [10 coeficientes]). En verde aparecen las tasas para los casos reconocidos correctamente. En rojo los casos clasificados erróneamente.

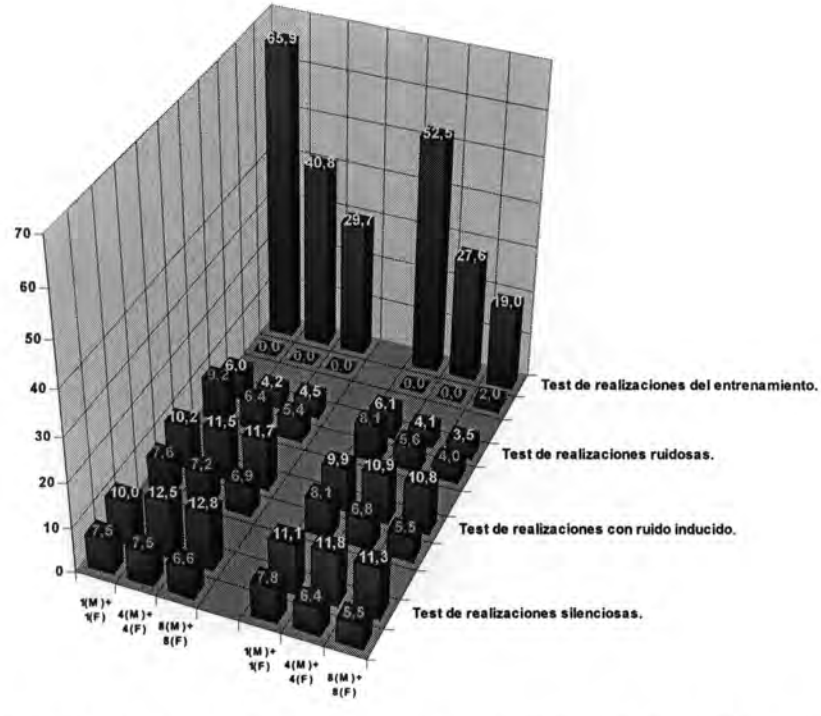

Figura A.44. Tasas de fiabilidad para el experimento LPC-11b (LPC-Bandas Energía Cepstrum [10 coeficientes]). En verde aparecen las tasas para los casos reconocidos correctamente. En rojo los casos clasificados erróneamente.

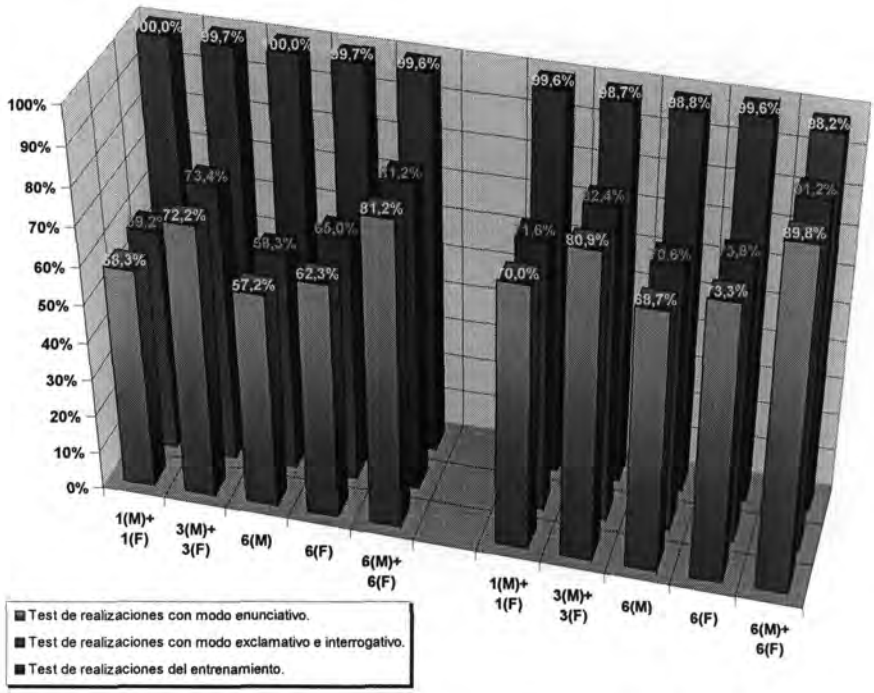

Figura A.45. Tasas de reconocimiento para el experimento LPC-12a (LPC-Bandas Energía Cepstrum [10 coeficientes + 10 coeficientes delta <5 tramas>]). La parte de la izquierda corresponde a los entrenamientos efectuados solamente con realizaciones con modo enunciativo.

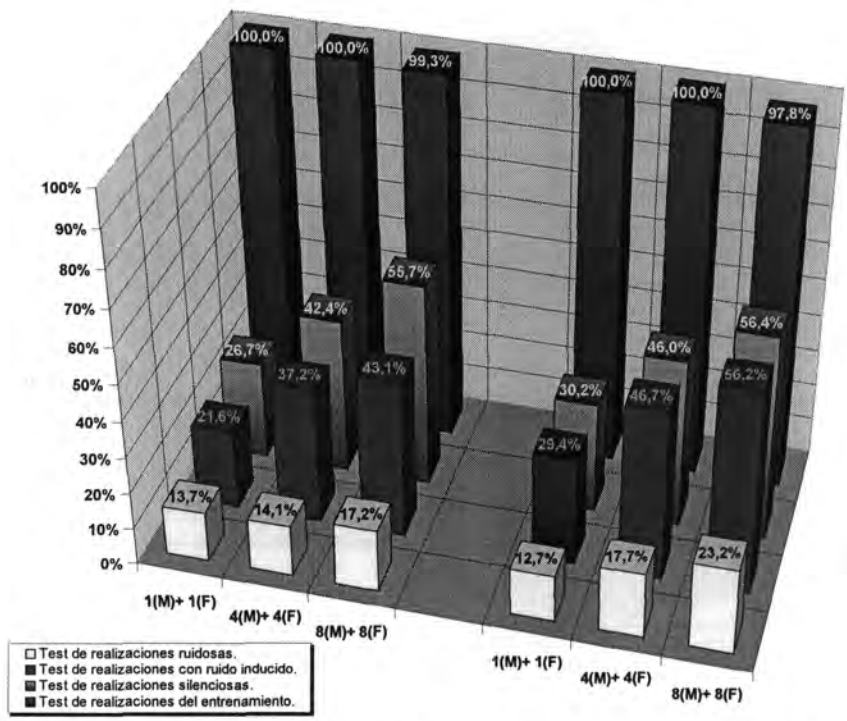

Figura A.46. Tasas de reconocimiento para el experimento LPC-12b (LPC-Bandas Energía Cepstrum [10 coeficientes + 10 coeficientes delta <5 tramas>]). La parte de la izquierda corresponde a los entrenamientos efectuados solamente con realizaciones silenciosas.

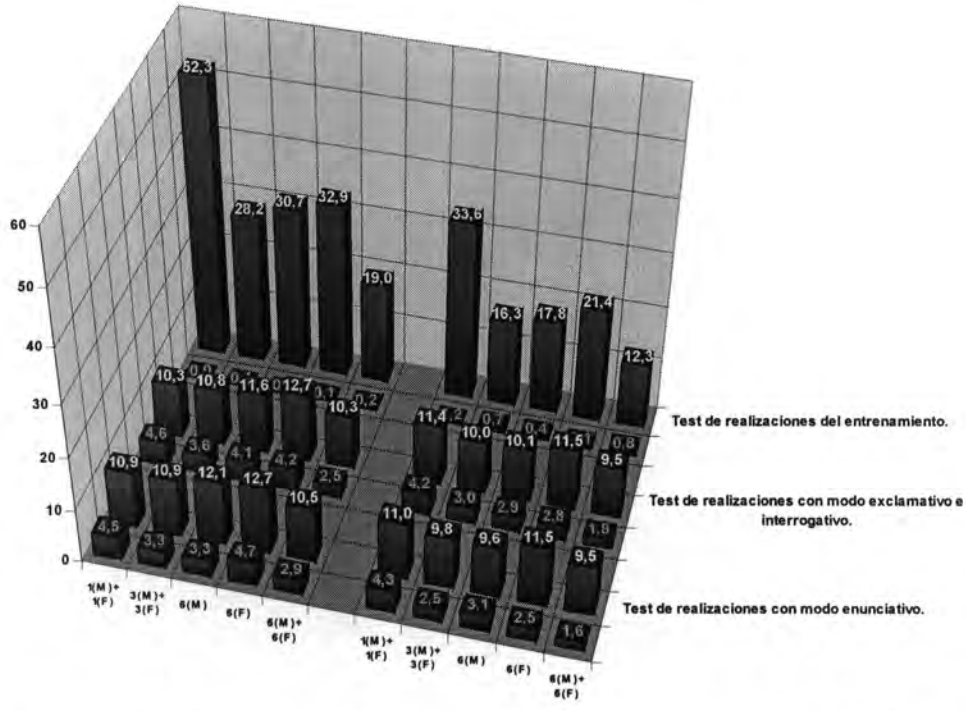

Figura A.47. Tasas de fiabilidad para el experimento LPC-12a (LPC-Bandas Energía Cepstrum [10 coeficientes + 10 coeficientes delta <5 tramas>]). En verde aparecen las tasas para los casos reconocidos correctamente. En rojo los casos clasificados erróneamente.

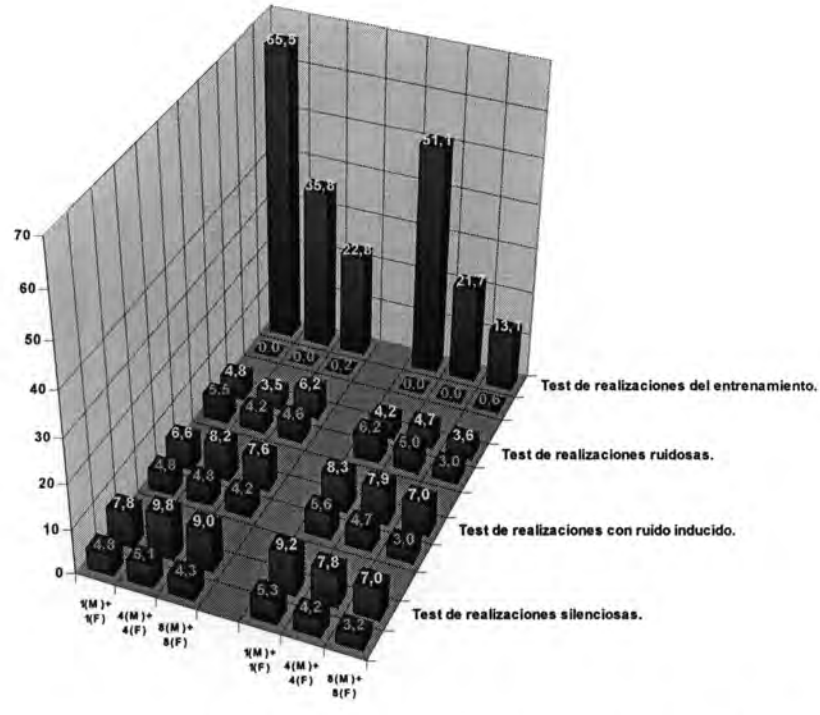

Figura A.48. Tasas de fiabilidad para el experimento LPC-12b (LPC-Bandas Energía Cepstrum [10 coeficientes + 10 coeficientes delta <5 tramas>]). En verde aparecen las tasas para los casos reconocidos correctamente. En rojo los casos clasificados erróneamente.

A.2. EXPERIMENTOS BASADOS EN FFT

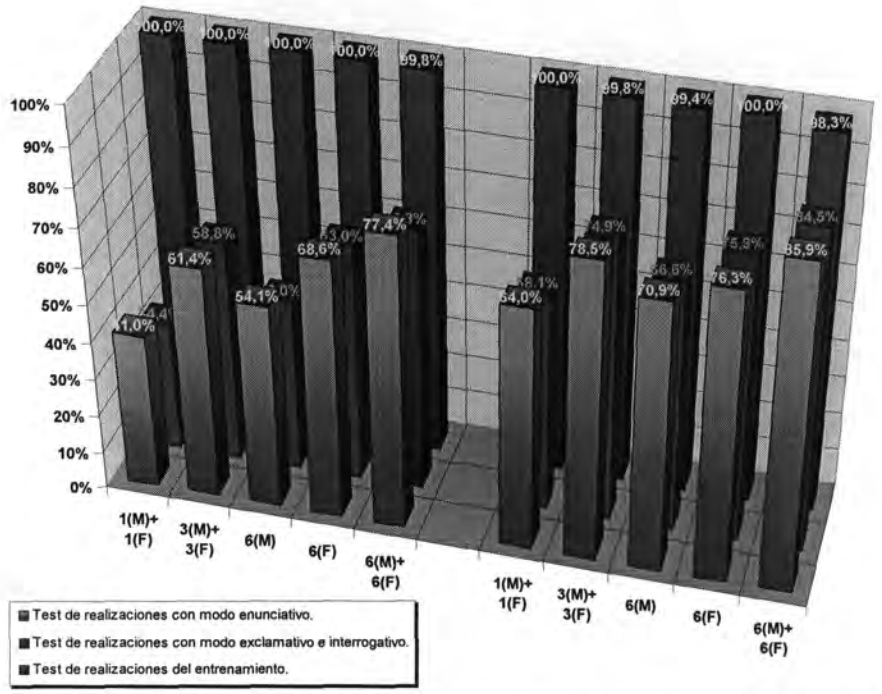

Figura A.49. Tasas de reconocimiento para el experimento FFT-1a (FFT-Bandas Energía [20 bandas]). La parte de la izquierda corresponde a los entrenamientos efectuados solamente con realizaciones con modo enunciativo.

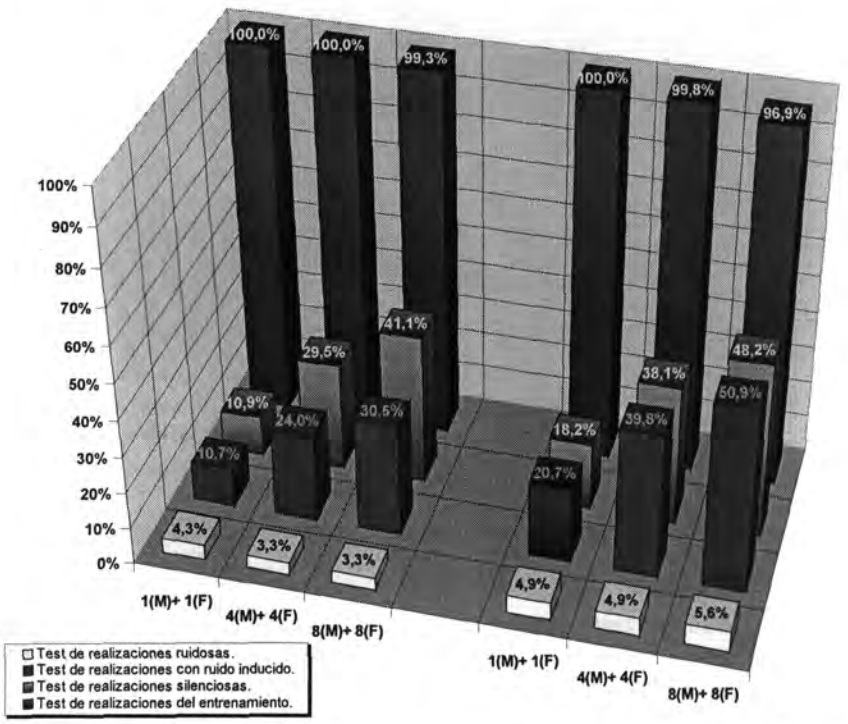

Figura A.50. Tasas de reconocimiento para el experimento FFT-1b (FFT-Bandas Energía [20 bandas]). La parte de la izquierda corresponde a los entrenamientos efectuados solamente con realizaciones silenciosas.

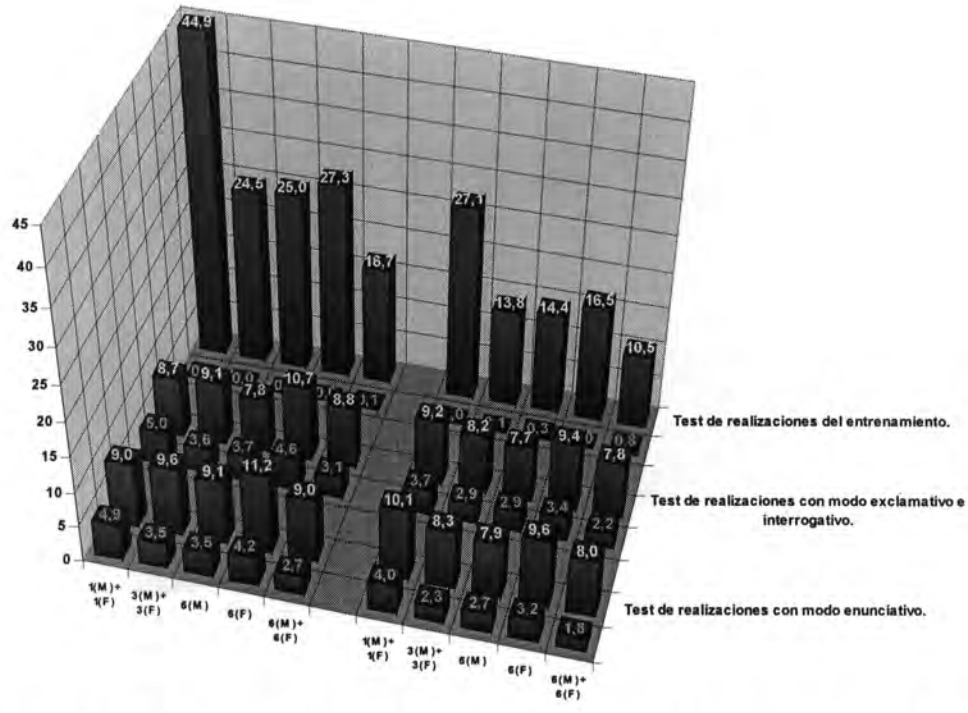

Figura A.51. Tasas de fiabilidad para el experimento FFT-1a (FFT-Bandas Energía [20 bandas]). En verde aparecen las tasas para los casos reconocidos correctamente. En rojo los casos clasificados erróneamente.

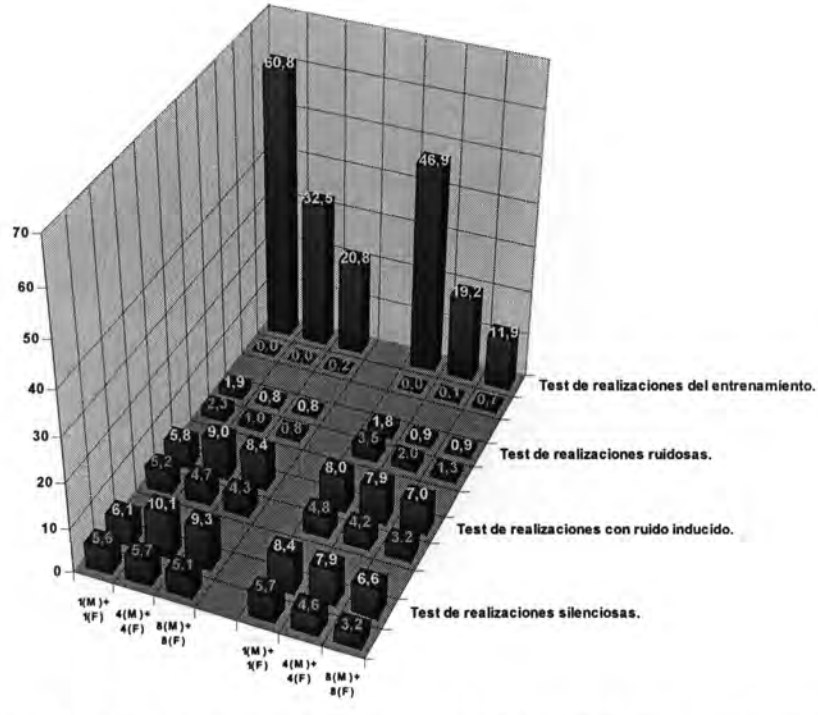

Figura A.52. Tasas de fiabilidad para el experimento FFT-1b (FFT-Bandas Energía [20 bandas]). En verde aparecen las tasas para los casos reconocidos correctamente. En rojo los casos clasificados erróneamente.

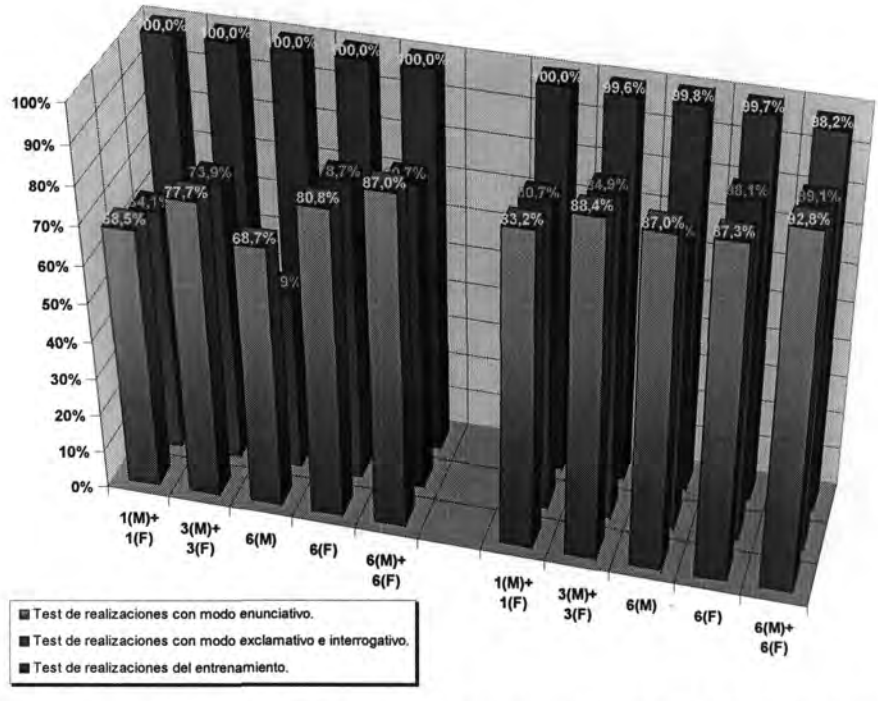

Figura A.53. Tasas de reconocimiento para el experimento FFT-2a (FFT-Bandas Energía [20 bandas + 20 bandas delta <5 tramas>]). La parte de la izquierda corresponde a los entrenamientos efectuados solamente con realizaciones con modo enunciativo.

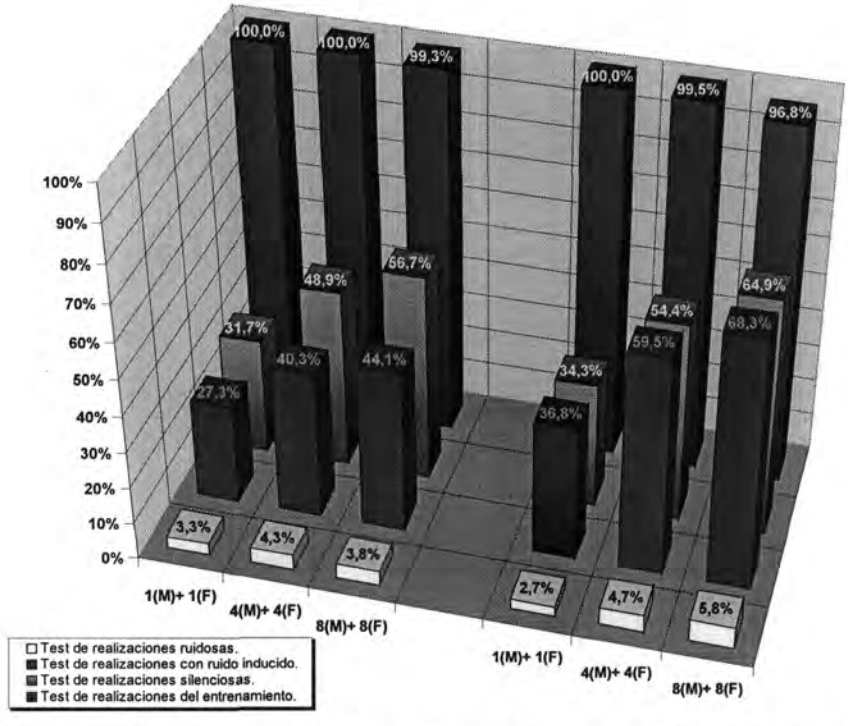

Figura A.54. Tasas de reconocimiento para el experimento FFT-2b (FFT-Bandas Energía [20 bandas + 20 bandas delta <5 tramas>]). La parte de la izquierda corresponde a los entrenamientos efectuados solamente con realizaciones silenciosas.

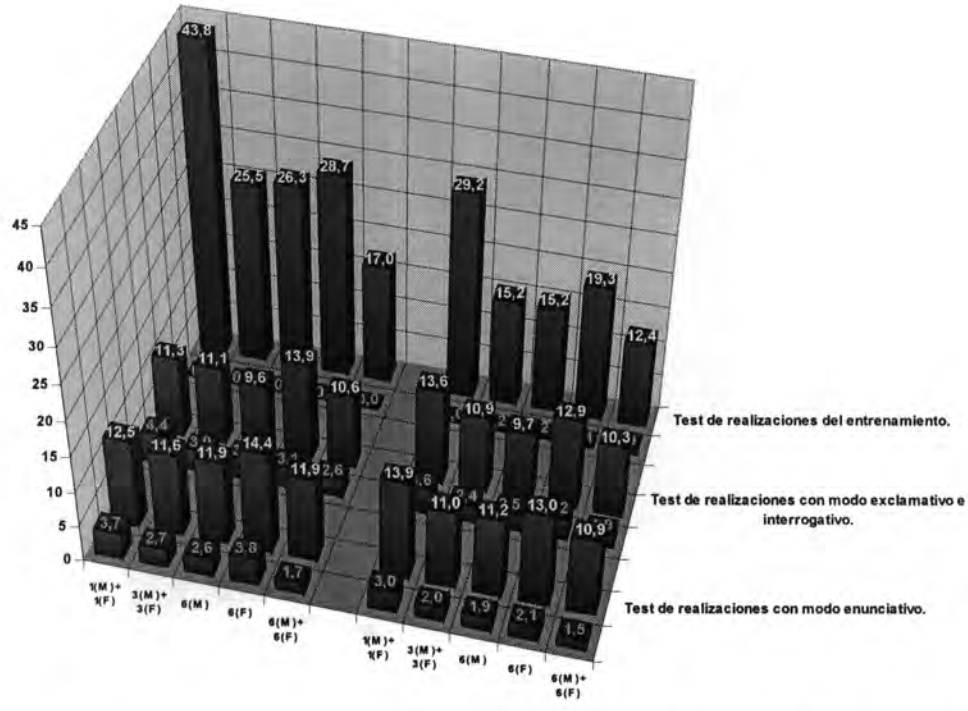

Figura A.55. Tasas de fiabilidad para el experimento FFT-2a (FFT-Bandas Energía [20 bandas + 20 bandas delta <5 tramas>]). En verde aparecen las tasas para los casos reconocidos correctamente. En rojo los casos clasificados erróneamente.

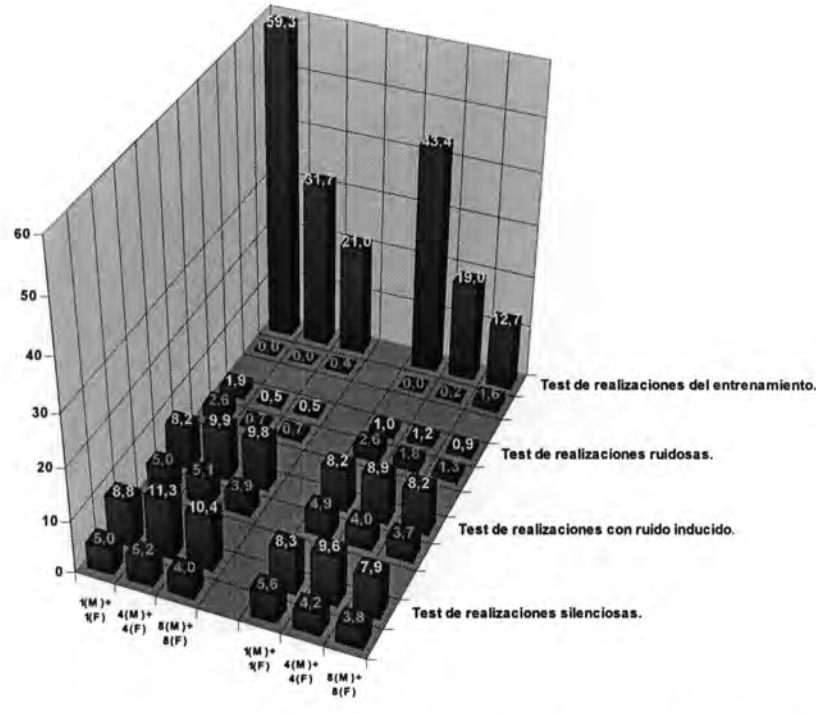

Figura A.56. Tasas de fiabilidad para el experimento FFT-2b (FFT-Bandas Energía [20 bandas + 20 bandas delta <5 tramas>]). En verde aparecen las tasas para los casos reconocidos correctamente. En rojo los casos clasificados erróneamente.

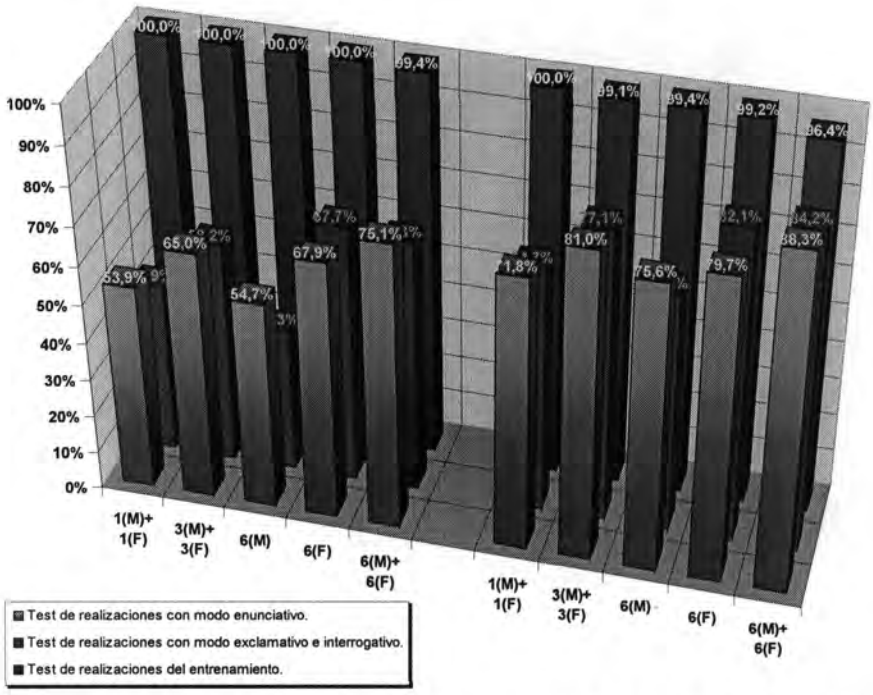

Figura A.57. Tasas de reconocimiento para el experimento FFT-3a (FFT-Bandas Energía [20 bandas + 20 bandas diferencia <2 tramas>]). La parte de la izquierda corresponde a los entrenamientos efectuados solamente con realizaciones con modo enunciativo.

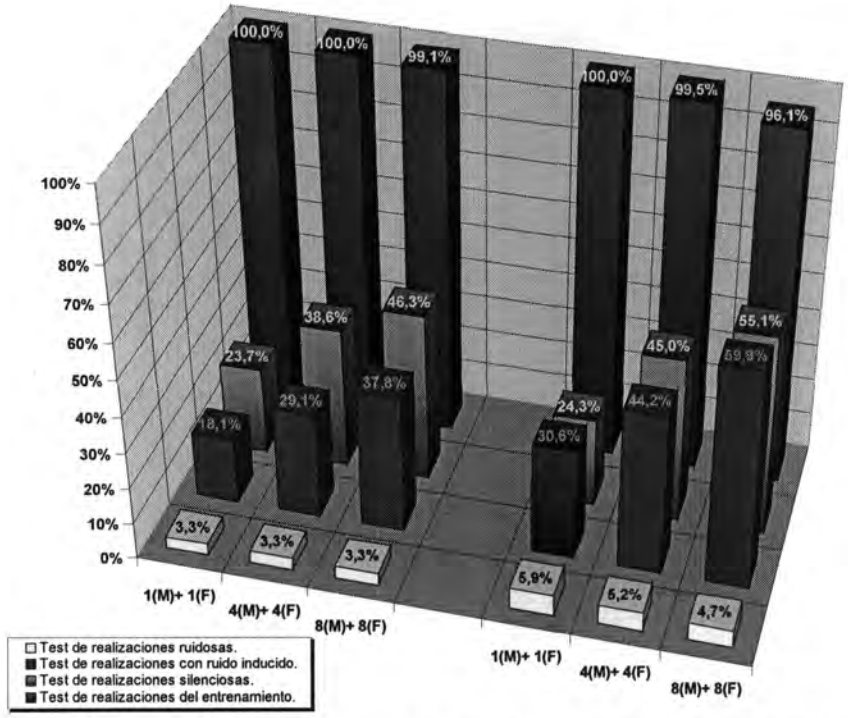

Figura A.58. Tasas de reconocimiento para el experimento FFT-3b (FFT-Bandas Energía [20 bandas + 20 bandas diferencia <2 tramas>]). La parte de la izquierda corresponde a los entrenamientos efectuados solamente con realizaciones silenciosas.

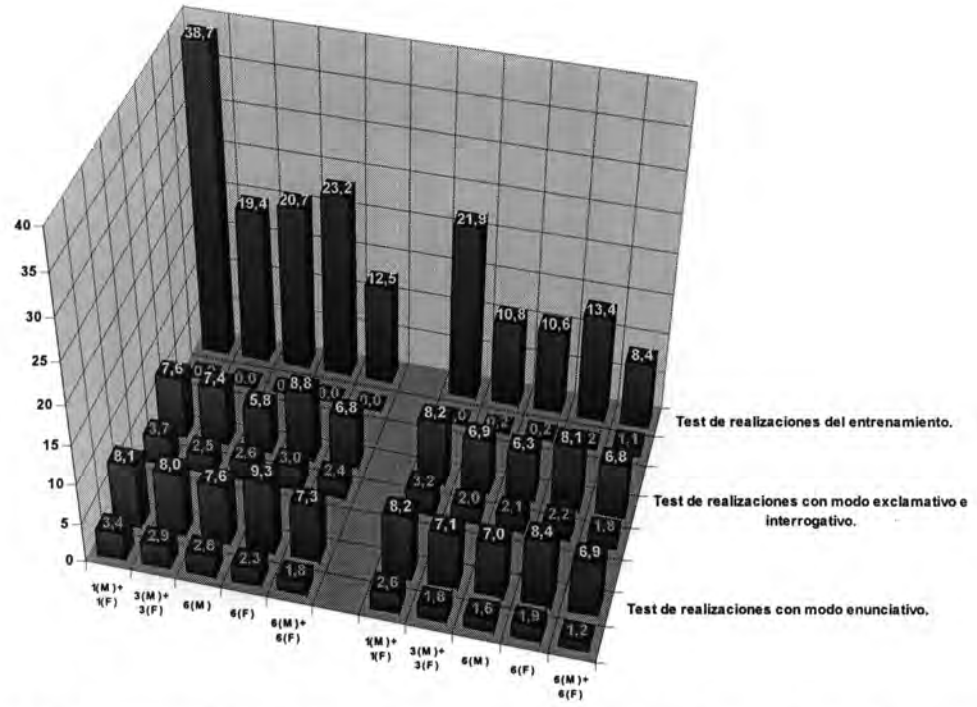

Figura A.59. Tasas de fiabilidad para el experimento FFT-3a (FFT-Bandas Energía [20 bandas + 20 bandas diferencia <2 tramas>]). En verde aparecen las tasas para los casos reconocidos correctamente. En rojo los casos clasificados erróneamente.

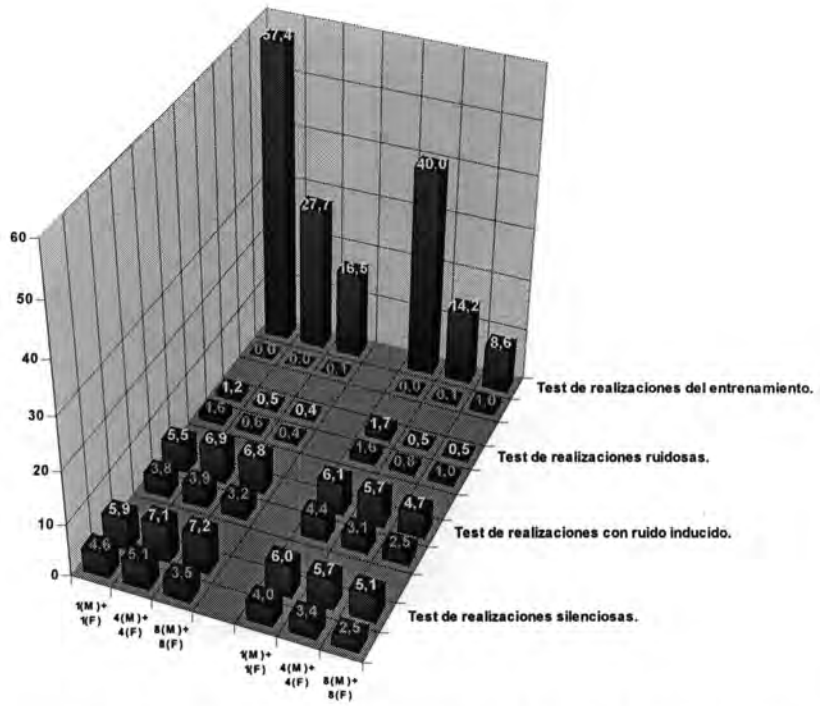

Figura A.60. Tasas de fiabilidad para el experimento FFT-3b (FFT-Bandas Energía [20 bandas + 20 bandas diferencia <2 tramas>]). En verde aparecen las tasas para los casos reconocidos correctamente. En rojo los casos clasificados erróneamente.

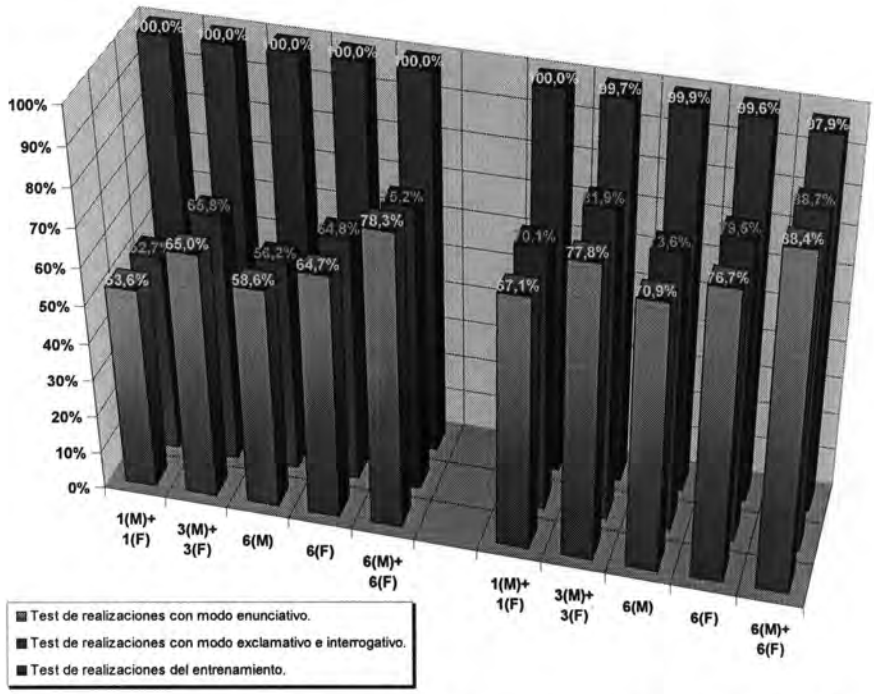

Figura A.61. Tasas de reconocimiento para el experimento FFT-4a (FFT-Bandas Energía Normalizadas [20 bandas]). La parte de la izquierda corresponde a los entrenamientos efectuados solamente con realizaciones con modo enunciativo.

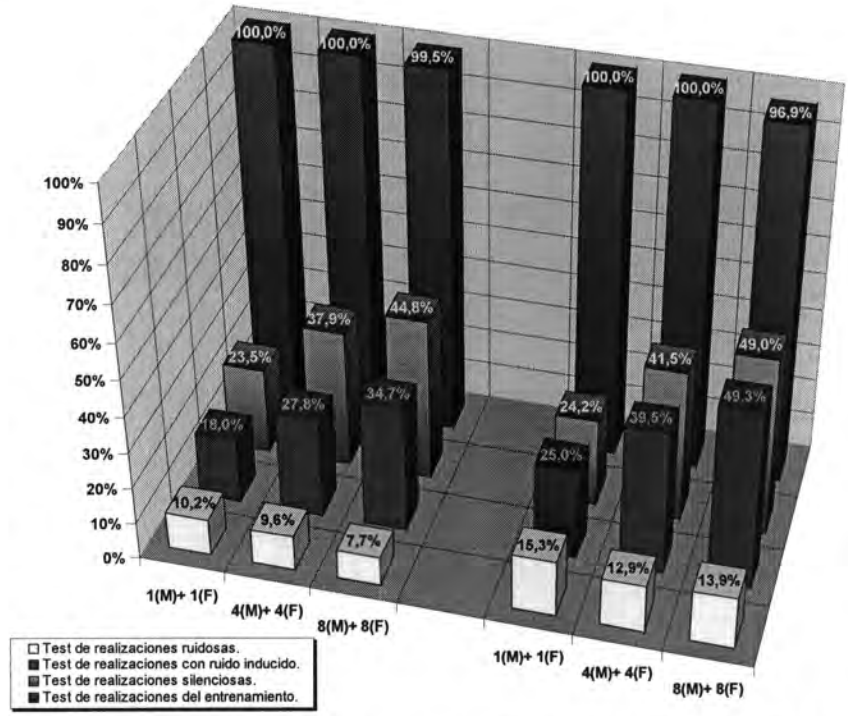

Figura A.62. Tasas de reconocimiento para el experimento FFT-4b (FFT-Bandas Energía Normalizadas [20 bandas]). La parte de la izquierda corresponde a los entrenamientos efectuados solamente con realizaciones silenciosas.

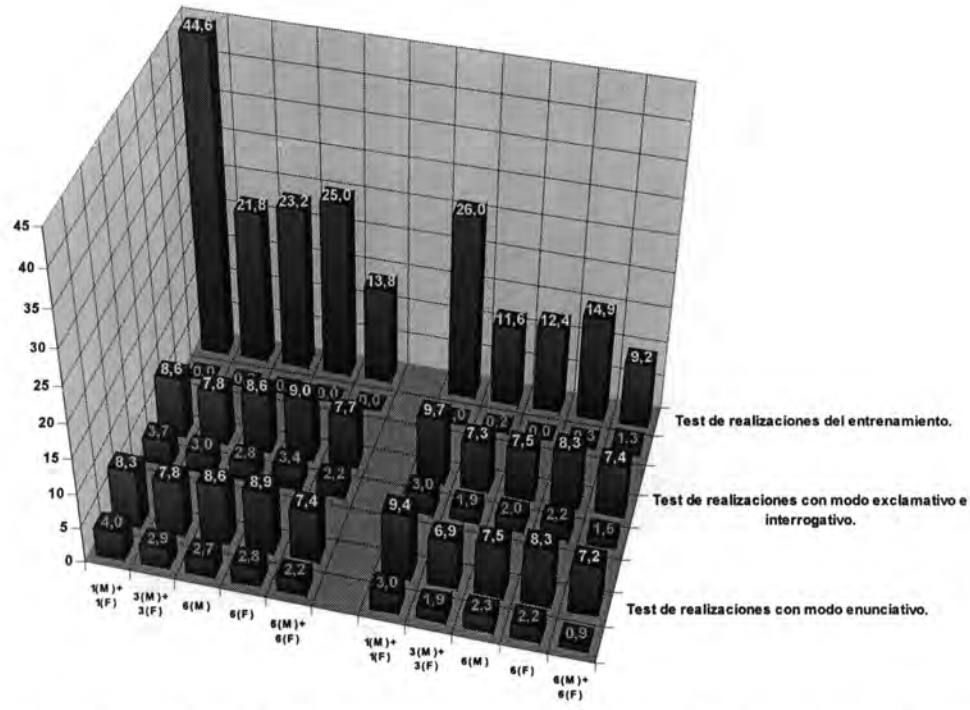

Figura A.63. Tasas de fiabilidad para el experimento FFT-4a (FFT-Bandas Energía Normalizadas [20 bandas]). En verde aparecen las tasas para los casos reconocidos correctamente. En rojo los casos clasificados erróneamente.

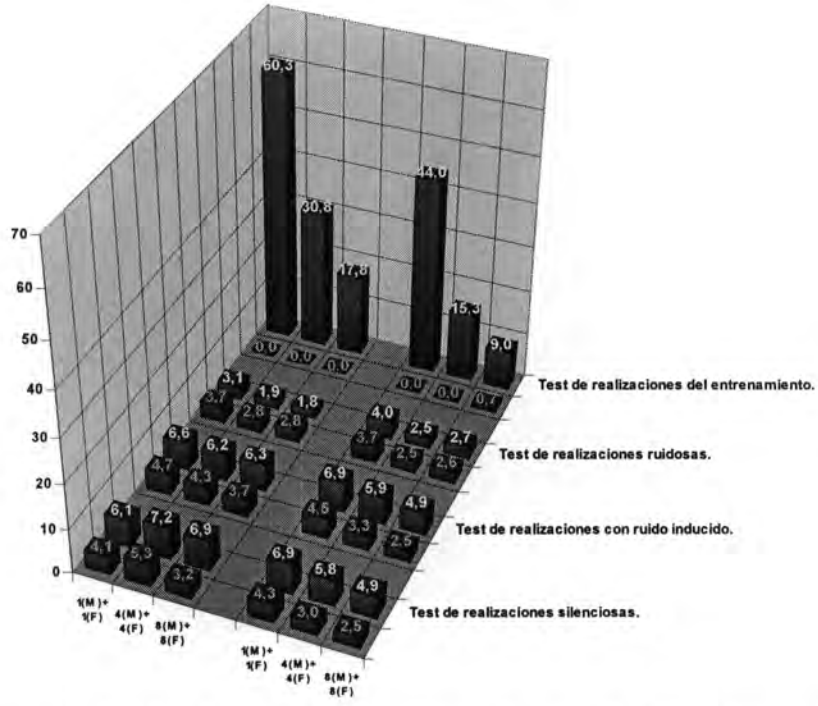

Figura A.64. Tasas de fiabilidad para el experimento FFT-4b (FFT-Bandas Energía Normalizadas [20 bandas]). En verde aparecen las tasas para los casos reconocidos correctamente. En rojo los casos clasificados erróneamente.

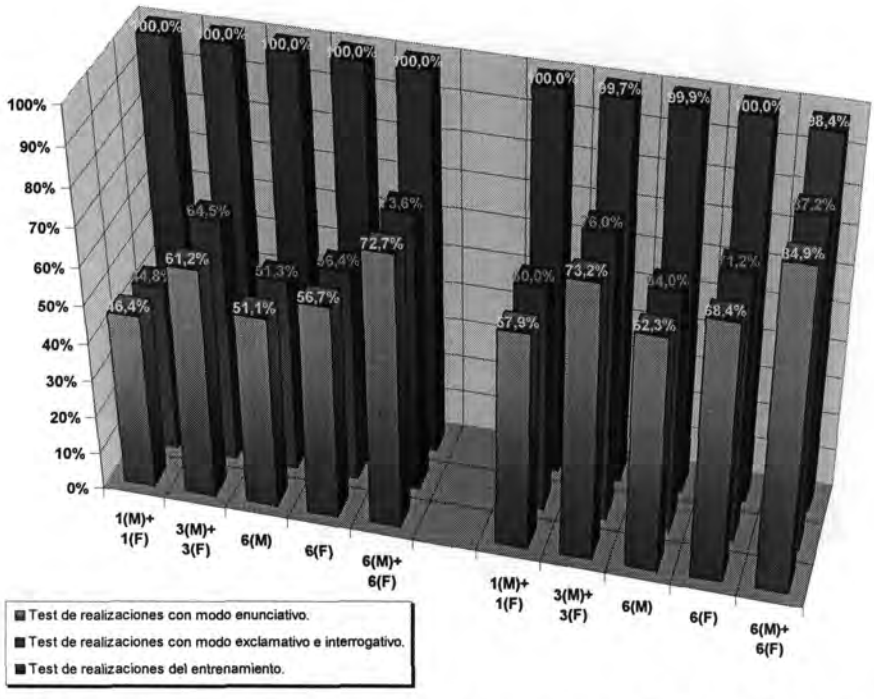

Figura A.65. Tasas de reconocimiento para el experimento FFT-5a (FFT-Bandas Energía Normalizadas [20 bandas + 20 bandas delta <5 tramas>]). La parte de la izquierda corresponde a los entrenamientos efectuados solamente con realizaciones con modo enunciativo.

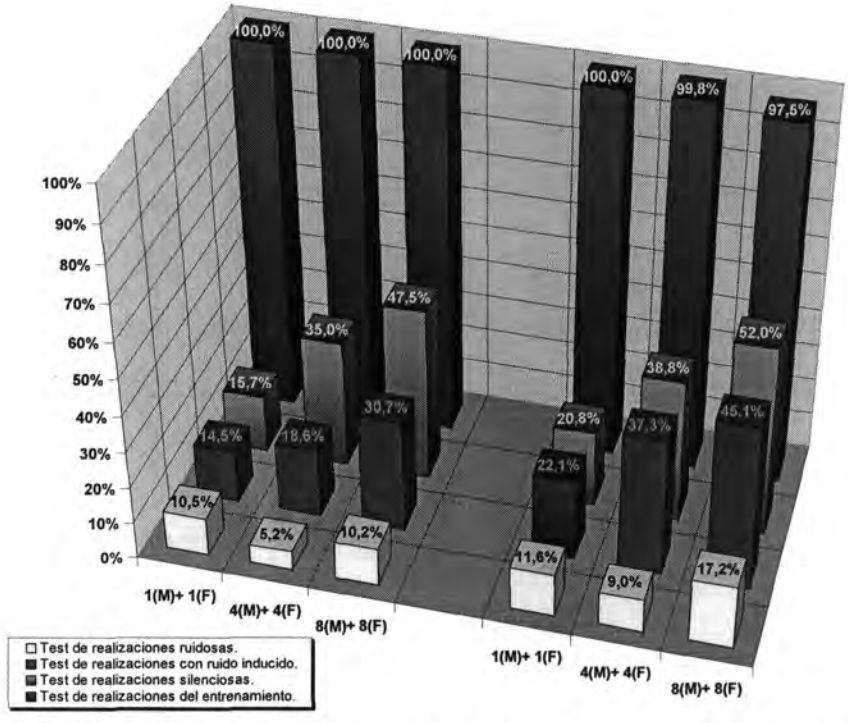

Figura A.66. Tasas de reconocimiento para el experimento FFT-5b (FFT-Bandas Energía Normalizadas [20 bandas + 20 bandas delta <5 tramas>]). La parte de la izquierda corresponde a los entrenamientos efectuados solamente con realizaciones silenciosas.

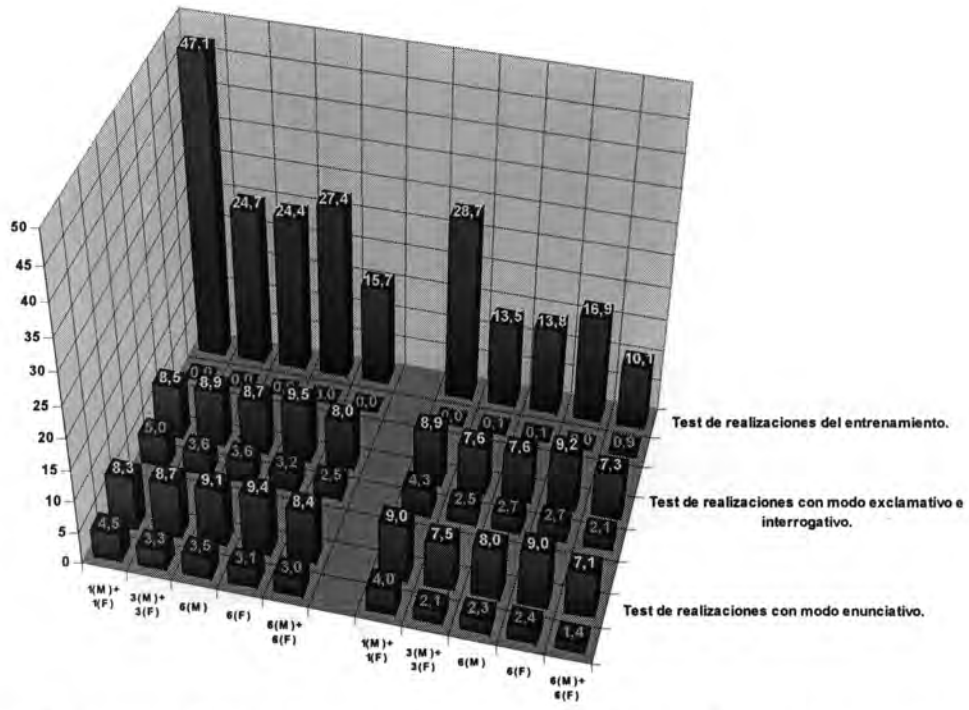

Figura A.67. Tasas de fiabilidad para el experimento FFT-5a (FFT-Bandas Energía Normalizadas [20 bandas + 20 bandas delta <5 tramas>]). En verde aparecen las tasas para los casos reconocidos correctamente. En rojo los casos clasificados erróneamente.

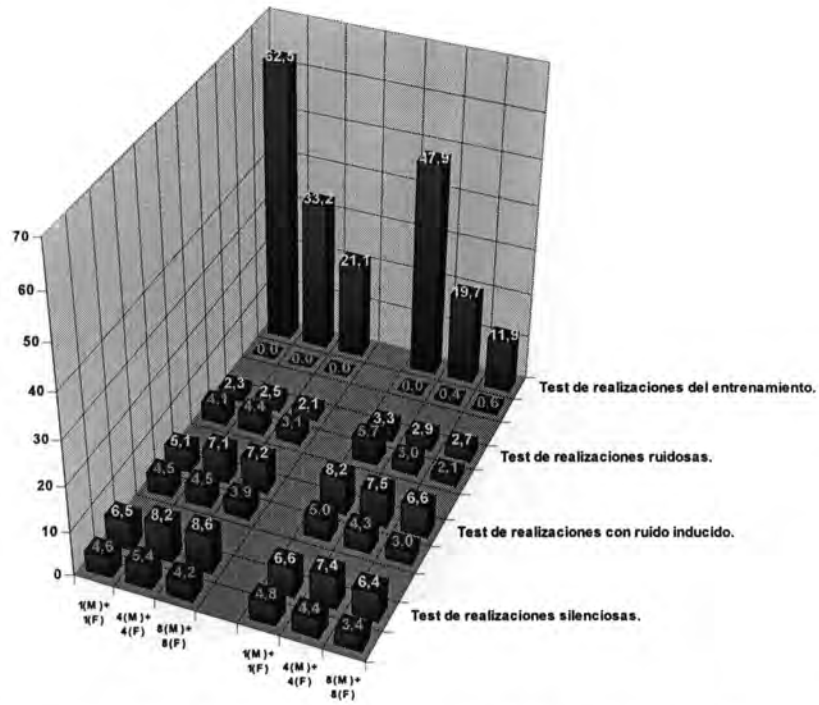

Figura A.68. Tasas de fiabilidad para el experimento FFT-5b (FFT-Bandas Energía Normalizadas [20 bandas + 20 bandas delta <5 tramas>]). En verde aparecen las tasas para los casos reconocidos correctamente. En rojo los casos clasificados erróneamente.

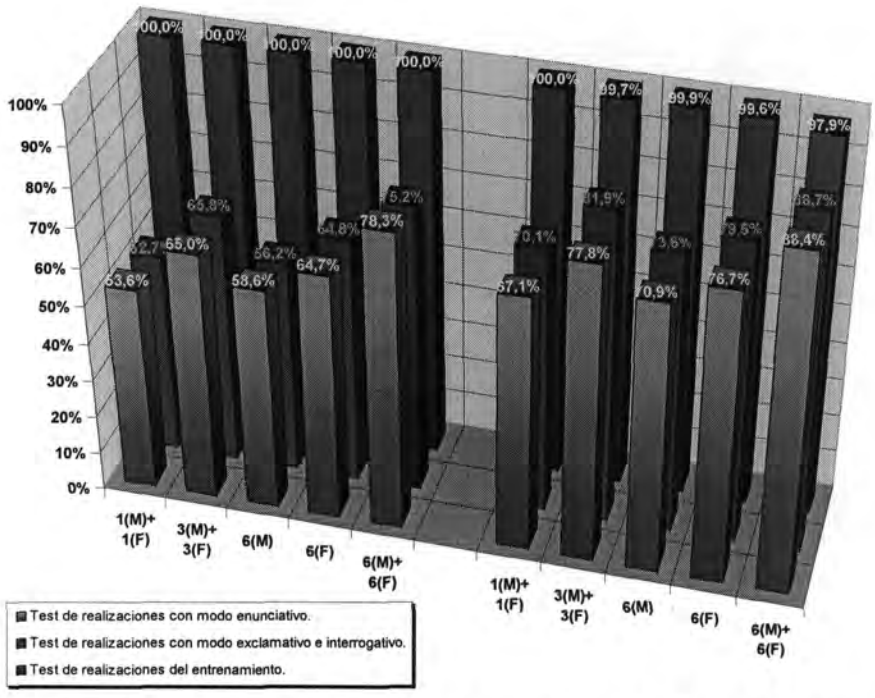

Figura A.69. Tasas de reconocimiento para el experimento FFT-6a (FFT-Bandas Energía Normalizadas [20 bandas + 20 bandas diferencia <2 tramas>]). La parte de la izquierda corresponde a los entrenamientos efectuados solamente con realizaciones con modo enunciativo.

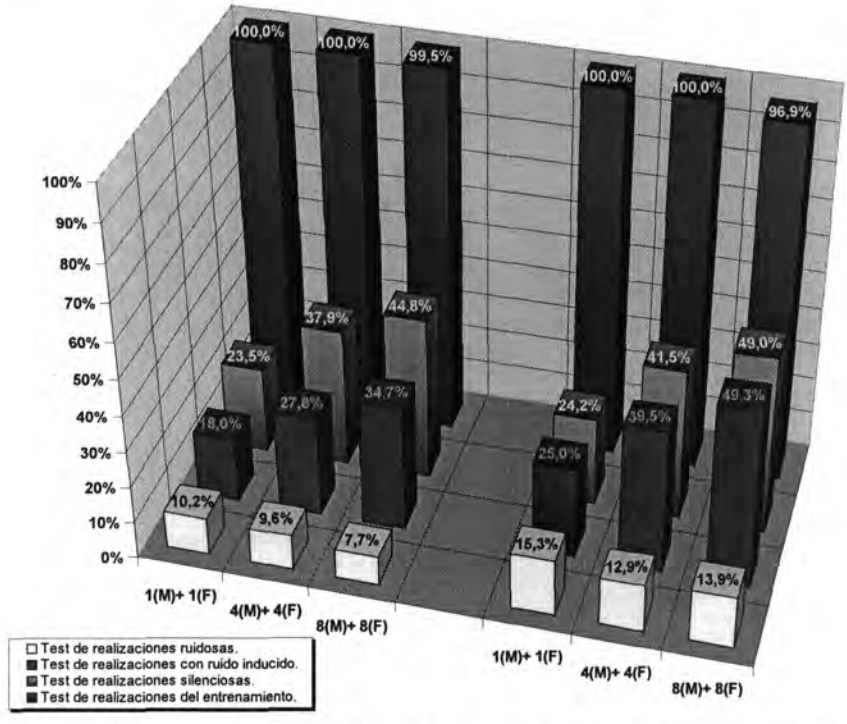

Figura A.70. Tasas de reconocimiento para el experimento FFT-6b (FFT-Bandas Energía Normalizadas [20 bandas + 20 bandas diferencia <2 tramas>]). La parte de la izquierda corresponde a los entrenamientos efectuados solamente con realizaciones silenciosas.

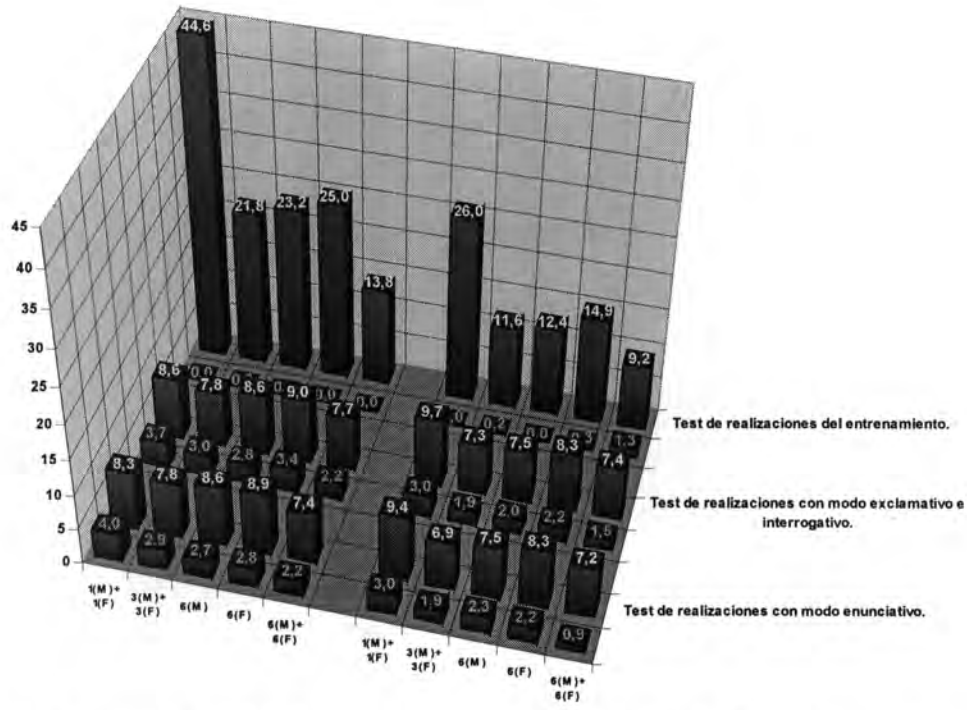

Figura A.71. Tasas de fiabilidad para el experimento FFT-6a (FFT-Bandas Energía Normalizadas [20 bandas + 20 bandas diferencia <2 tramas>]). En verde aparecen las tasas para los casos reconocidos correctamente. En rojo los casos clasificados erróneamente.

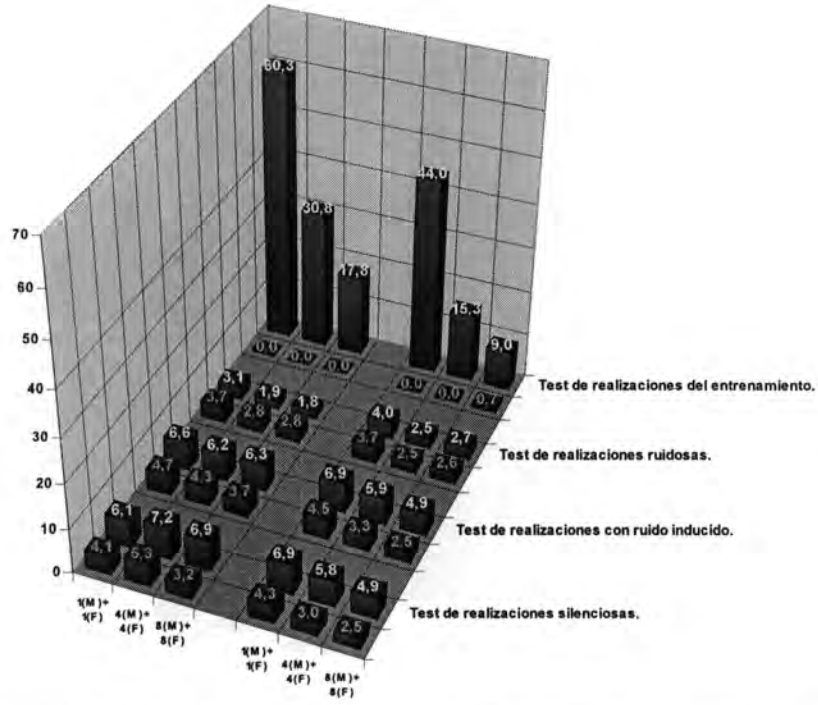

Figura A.72. Tasas de fiabilidad para el experimento FFT-6b (FFT-Bandas Energía Normalizadas [20 bandas + 20 bandas diferencia <2 tramas>]). En verde aparecen las tasas para los casos reconocidos correctamente. En rojo los casos clasificados erróneamente.

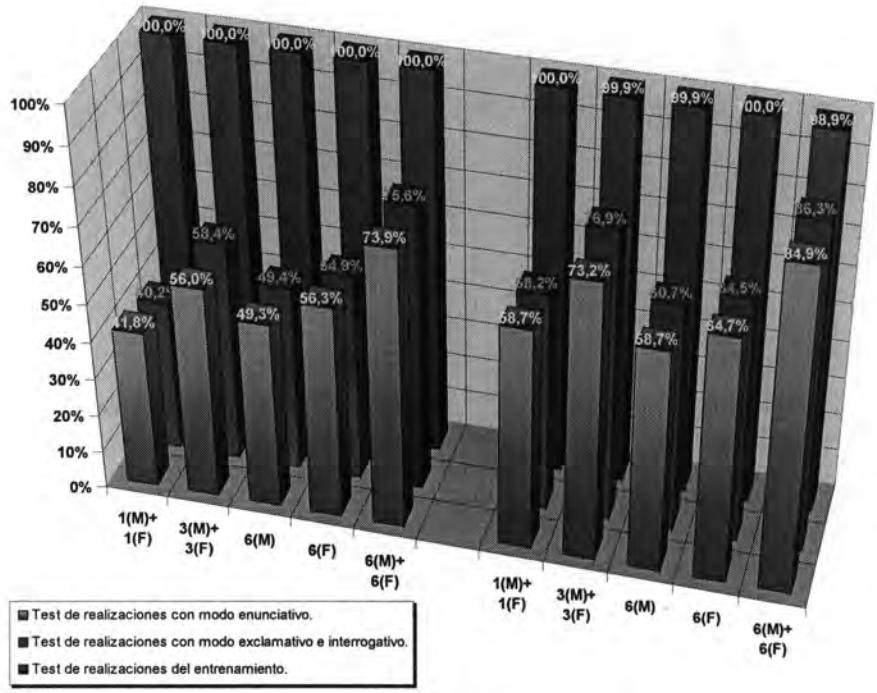

Figura A.73. Tasas de reconocimiento para el experimento FFT-7a (FFT-Cepstrum [10 coeficientes]). La parte de la izquierda corresponde a los entrenamientos efectuados solamente con realizaciones con modo enunciativo.

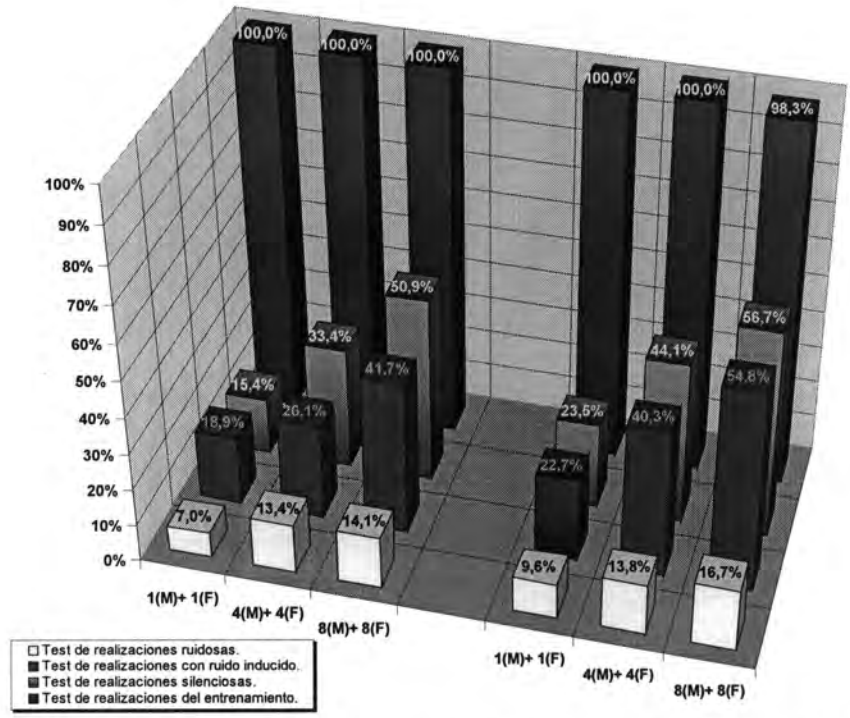

Figura A.74. Tasas de reconocimiento para el experimento FFT-7b (FFT-Cepstrum [10 coeficientes]). La parte de la izquierda corresponde a los entrenamientos efectuados solamente con realizaciones silenciosas.

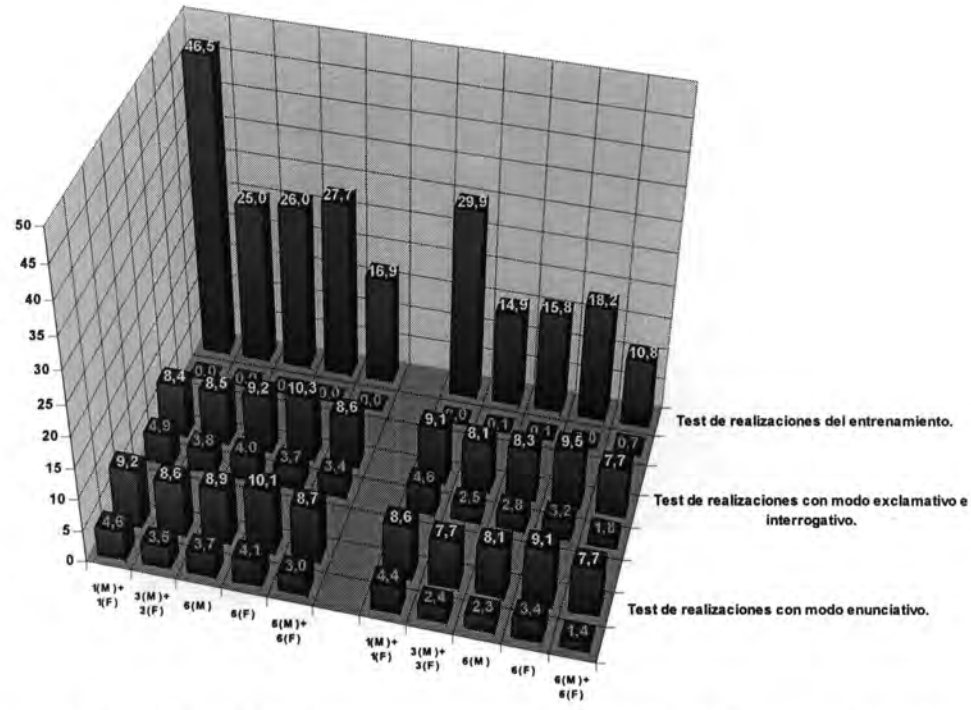

Figura A.75. Tasas de fiabilidad para el experimento FFT-7a (FFT-Cepstrum [10 coeficientes]). En verde aparecen las tasas para los casos reconocidos correctamente. En rojo los casos clasificados erróneamente.

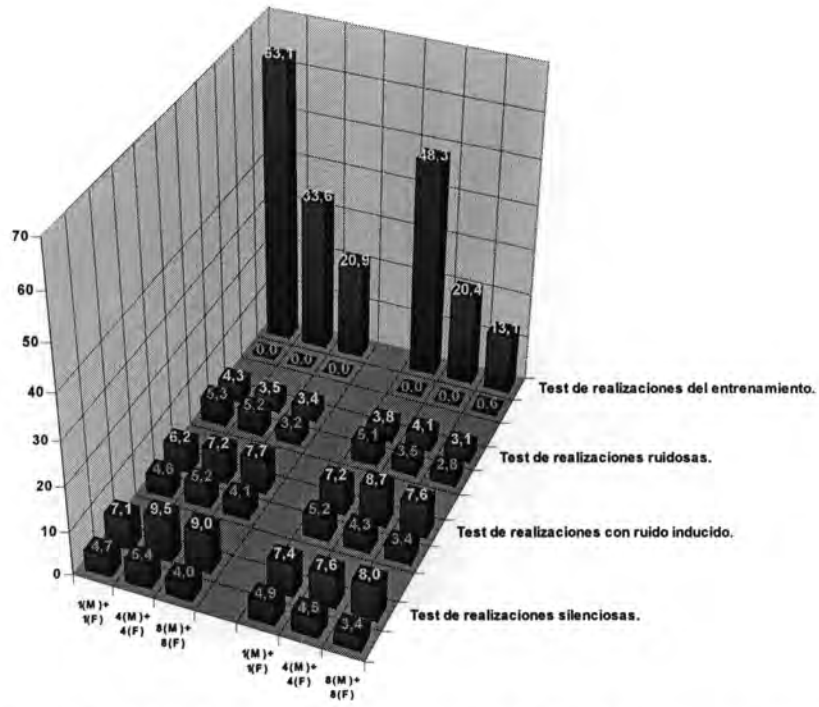

Figura A.76. Tasas de fiabilidad para el experimento FFT-7b (FFT-Cepstrum [10 coeficientes]). En verde aparecen las tasas para los casos reconocidos correctamente. En rojo los casos clasificados erróneamente.

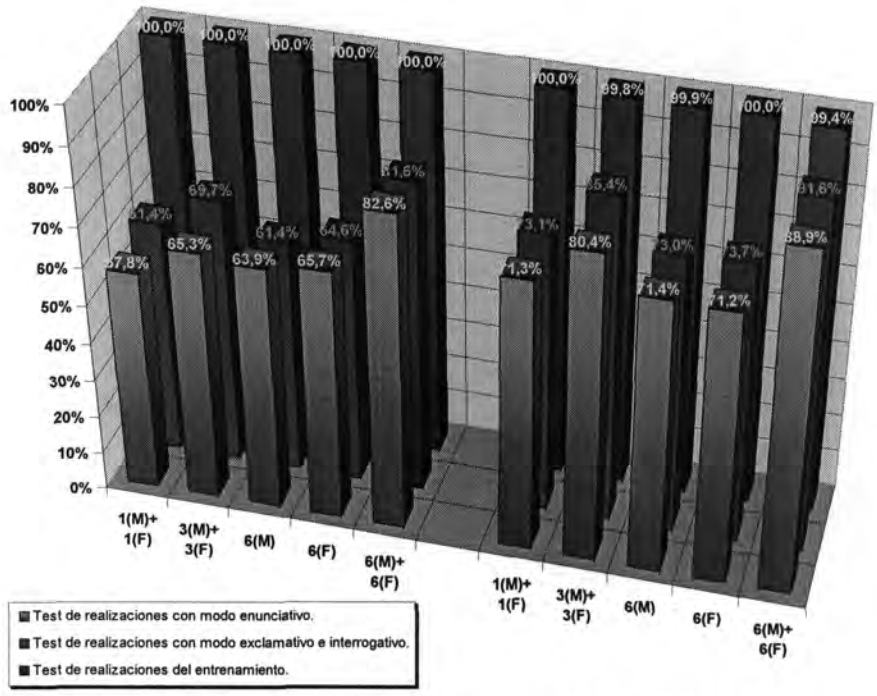

Figura A.77. Tasas de reconocimiento para el experimento FFT-8a (FFT-Cepstrum [10 coeficientes + 10 coeficientes delta <5 tramas>]). La parte de la izquierda corresponde a los entrenamientos efectuados solamente con realizaciones con modo enunciativo.

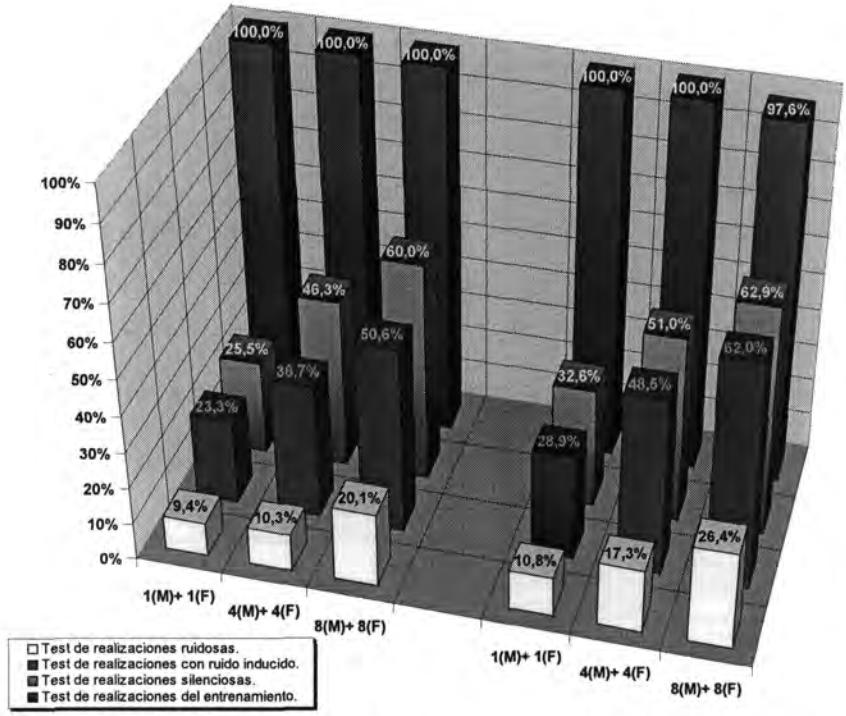

Figura A.78. Tasas de reconocimiento para el experimento FFT-8b (FFT-Cepstrum [10 coeficientes + 10 coeficientes delta <5 tramas>]). La parte de la izquierda corresponde a los entrenamientos efectuados solamente con realizaciones silenciosas.

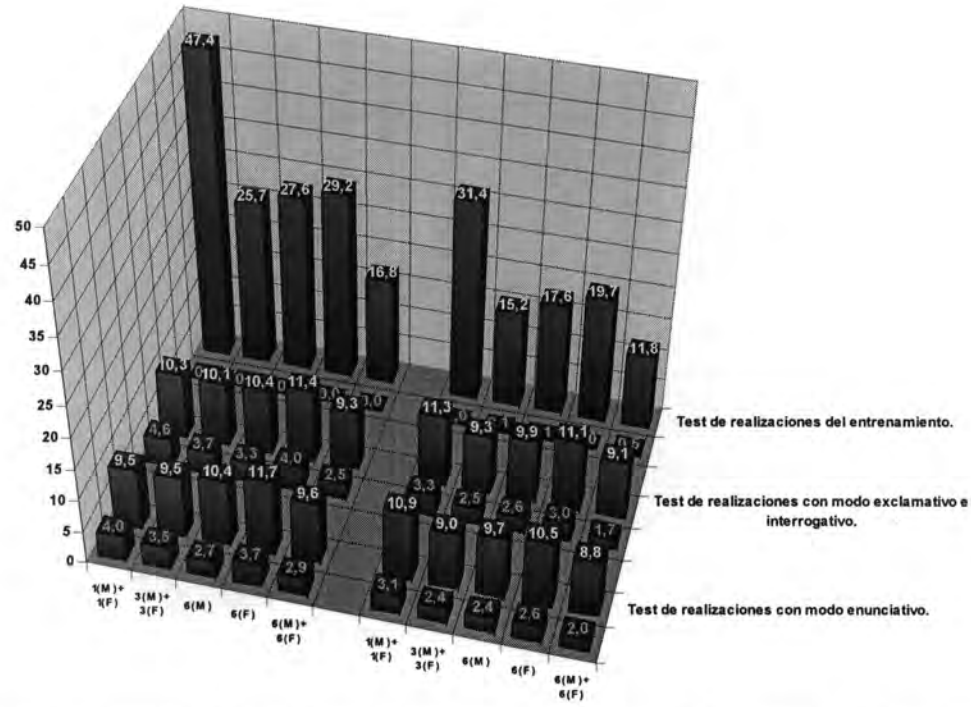

Figura A.79. Tasas de fiabilidad para el experimento FFT-8a (FFT-Cepstrum [10 coeficientes + 10 coeficientes delta <5 tramas>]). En verde aparecen las tasas para los casos reconocidos correctamente. En rojo los casos clasificados erróneamente.

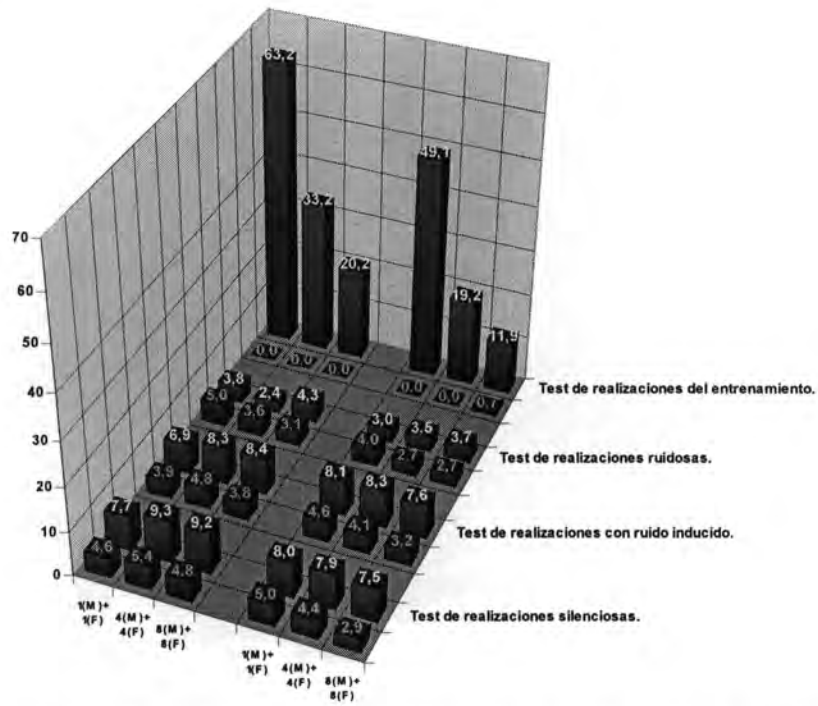

Figura A.80. Tasas de fiabilidad para el experimento FFT-8b (FFT-Cepstrum [10 coeficientes + 10 coeficientes delta <5 tramas>]). En verde aparecen las tasas para los casos reconocidos correctamente. En rojo los casos clasificados erróneamente.

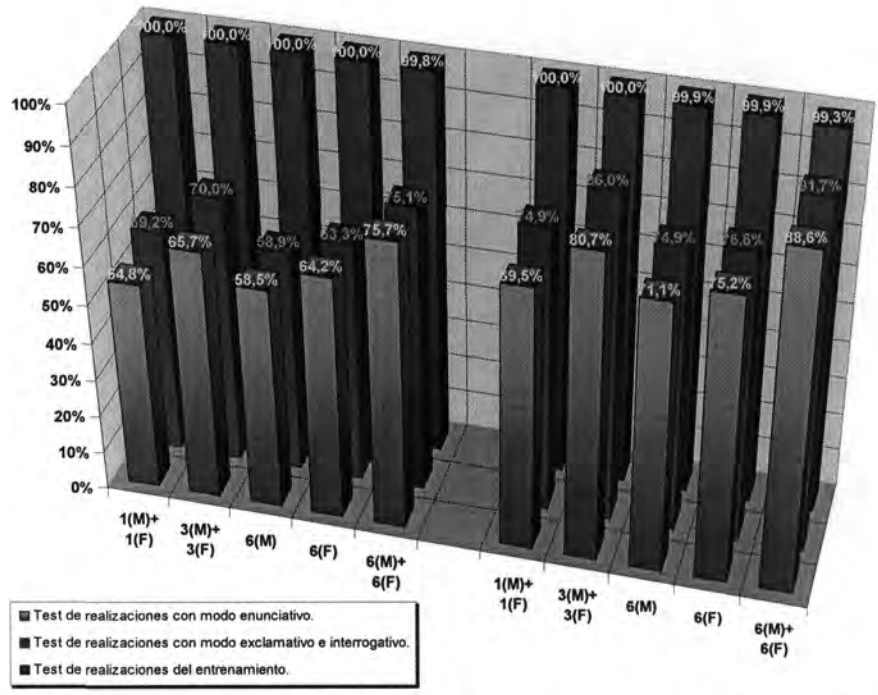

Figura A.81. Tasas de reconocimiento para el experimento FFT-9a (FFT-Cepstrum [10 coeficientes + 10 coeficientes delta <5 tramas> + 10 coeficientes delta-delta]). La parte de la izquierda corresponde a los entrenamientos efectuados solamente con realizaciones con modo enunciativo.

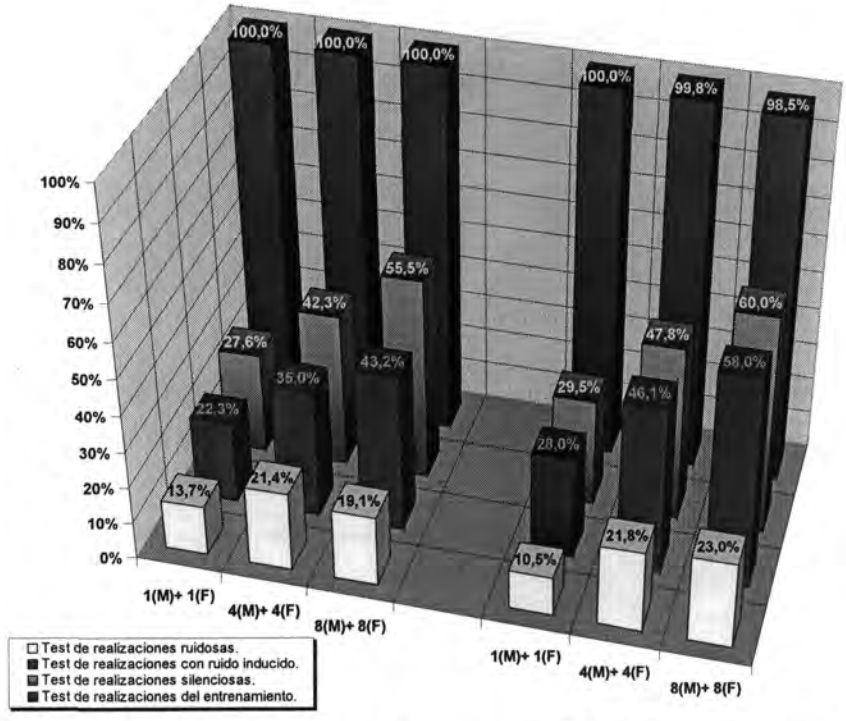

Figura A.82. Tasas de reconocimiento para el experimento FFT-9b (FFT-Cepstrum [10 coeficientes + 10 coeficientes delta <5 tramas> + 10 coeficientes delta-delta]). La parte de la izquierda corresponde a los entrenamientos efectuados solamente con realizaciones silenciosas.

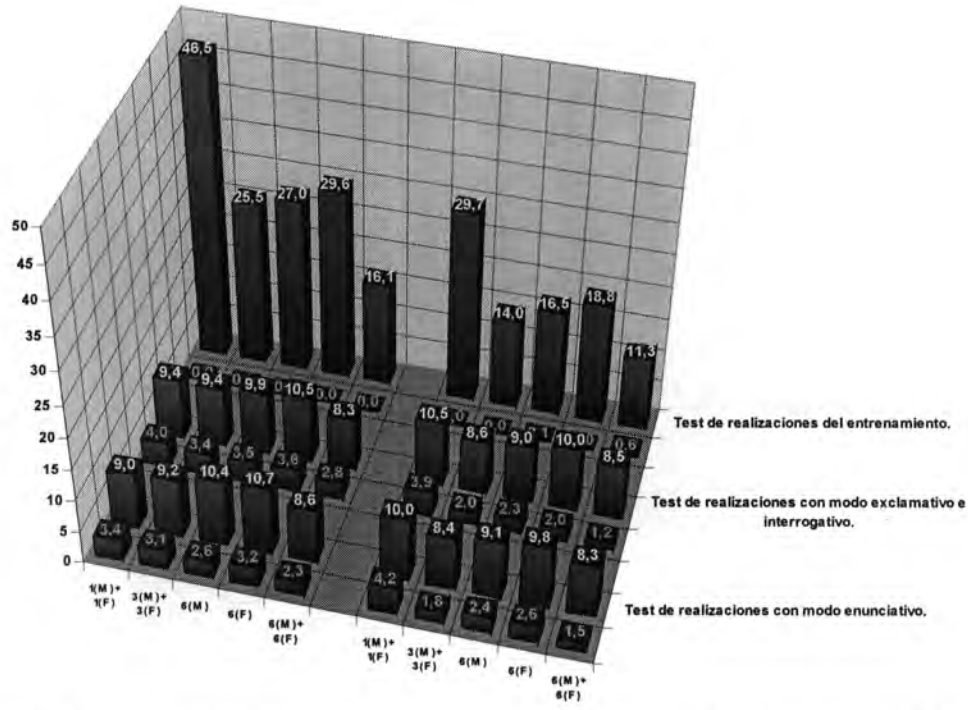

Figura A.83. Tasas de fiabilidad para el experimento FFT-9a (FFT-Cepstrum [10 coeficientes + 10 coeficientes delta <5 tramas> + 10 coeficientes delta-delta]). En verde aparecen las tasas para los casos reconocidos correctamente. En rojo los casos clasificados erróneamente.

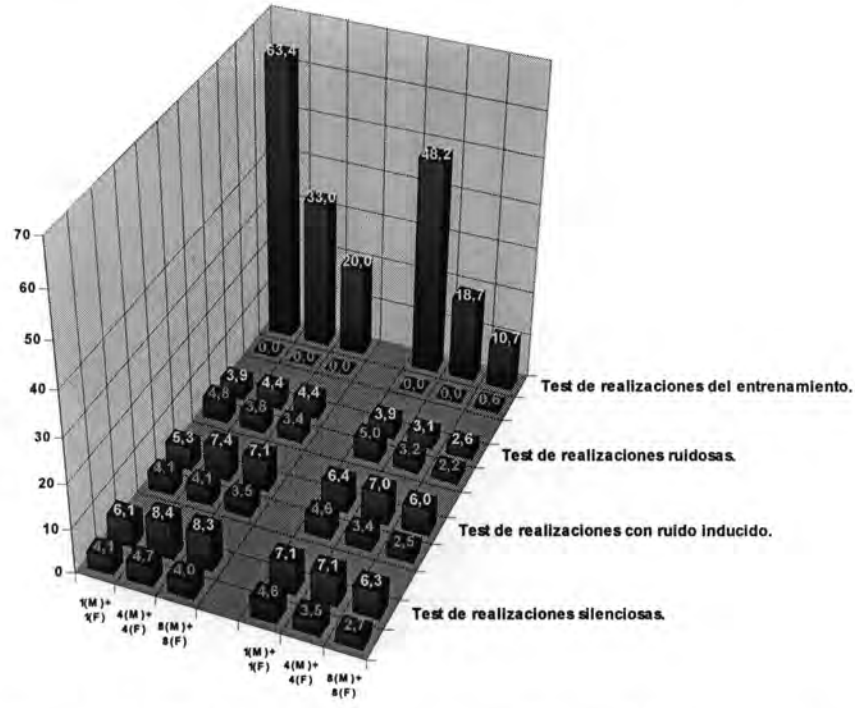

Figura A.84. Tasas de fiabilidad para el experimento FFT-9b (FFT-Cepstrum [10 coeficientes + 10 coeficientes delta <5 tramas> + 10 coeficientes delta-delta]). En verde aparecen las tasas para los casos reconocidos correctamente. En rojo los casos clasificados erróneamente.

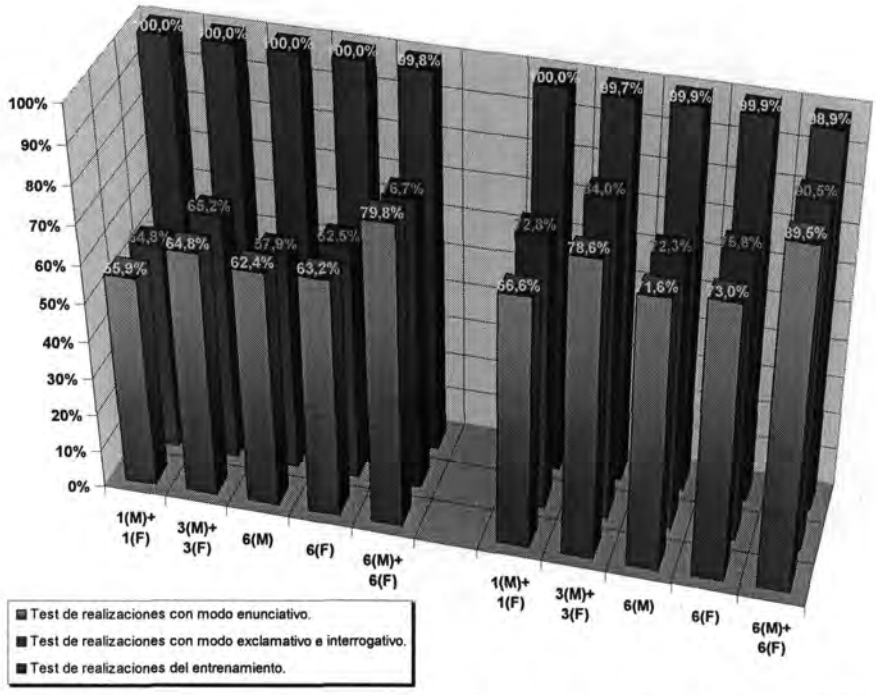

Figura A.85. Tasas de reconocimiento para el experimento FFT-10a (FFT-Cepstrum [10 coeficientes + 10 coeficientes delta <5 tramas> + 10 coeficientes delta-delta + energía trama + energía trama diferencia <2 tramas>]). La parte de la izquierda corresponde a los entrenamientos efectuados solamente con realizaciones con modo enunciativo.

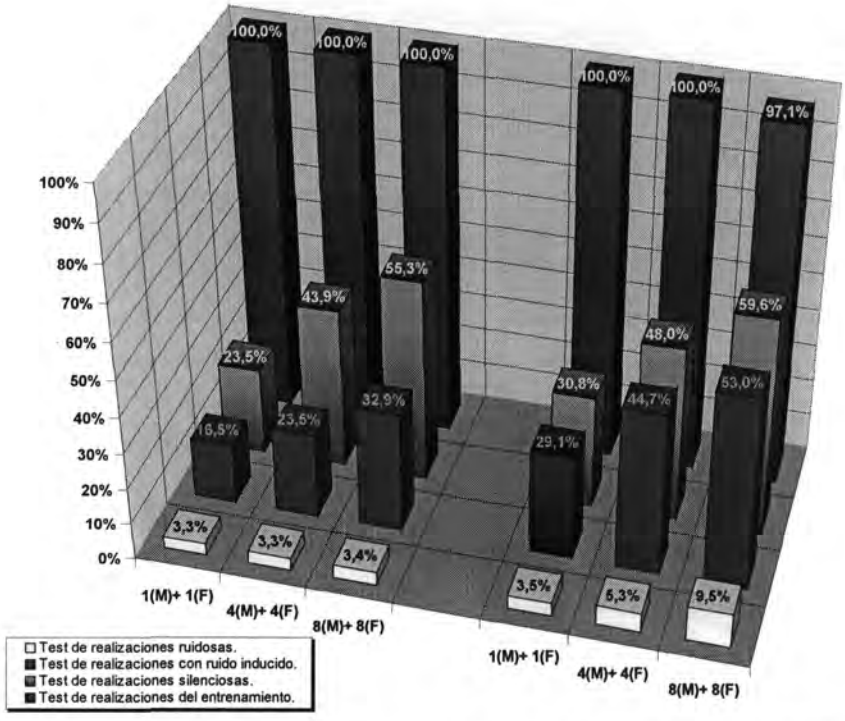

Figura A.86. Tasas de reconocimiento para el experimento FFT-10b (FFT-Cepstrum [10 coeficientes + 10 coeficientes delta <5 tramas> + 10 coeficientes delta-delta + energía trama + energía trama diferencia <2 tramas>]). La parte de la izquierda corresponde a los entrenamientos efectuados solamente con realizaciones silenciosas.

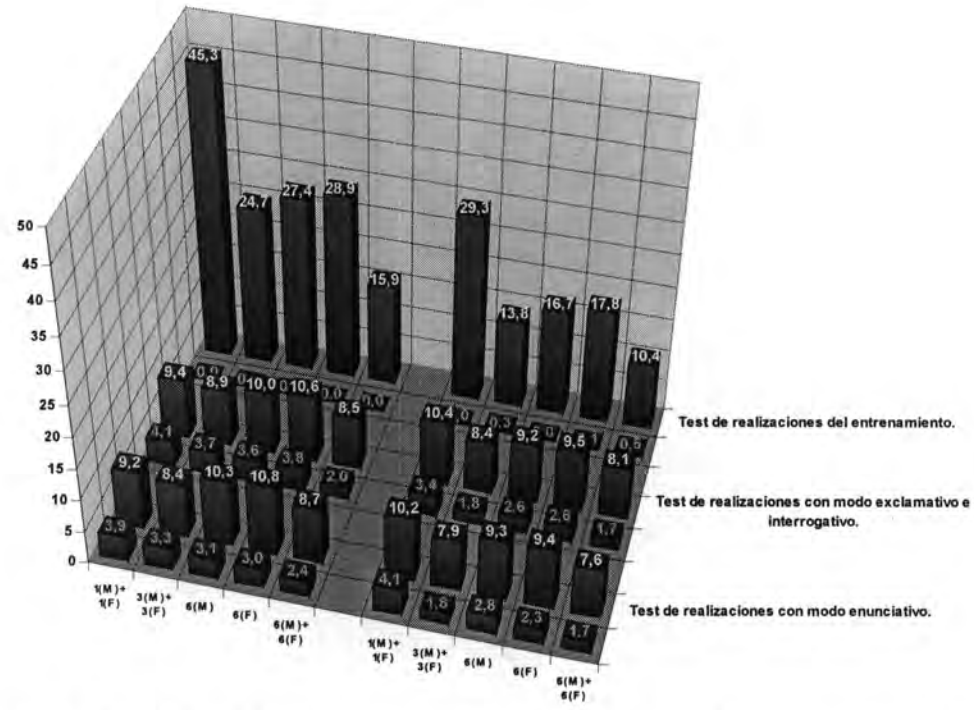

Figura A.87. Tasas de fiabilidad para el experimento FFT-10a (FFT-Cepstrum [10 coeficientes + 10 coeficientes delta <5 tramas> + 10 coeficientes delta-delta + energía trama + energía trama diferencia <2 tramas>]). En verde aparecen las tasas para los casos reconocidos correctamente. En rojo los casos clasificados erróneamente.

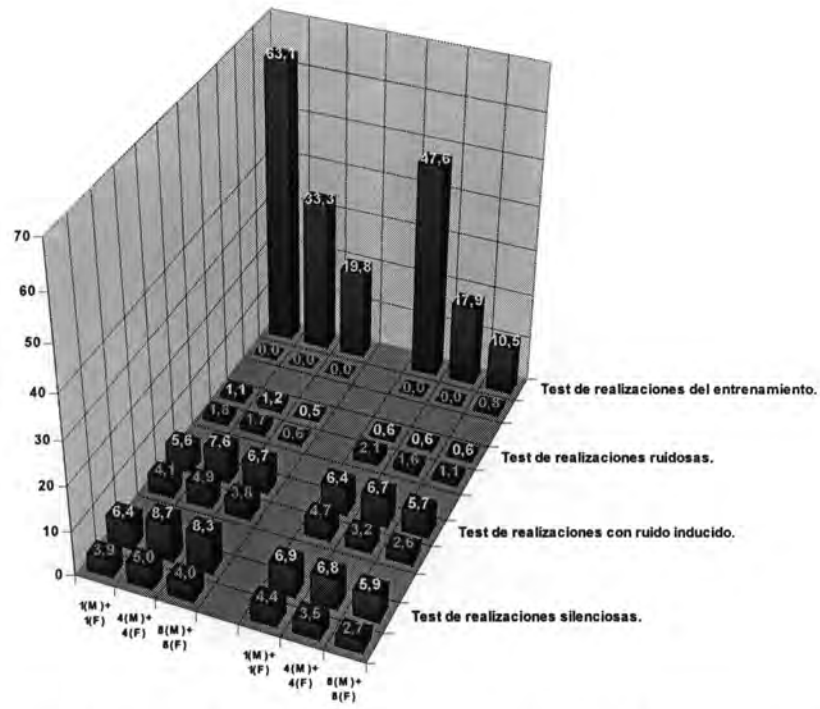

Figura A.88. Tasas de fiabilidad para el experimento FFT-10b (FFT-Cepstrum [10 coeficientes + 10 coeficientes delta <5 tramas> + 10 coeficientes delta-delta + energía trama + energía trama diferencia <2 tramas>]). En verde aparecen las tasas para los casos reconocidos correctamente. En rojo los casos clasificados erróneamente.

A.3. EXPERIMENTOS BASADOS EN PLP

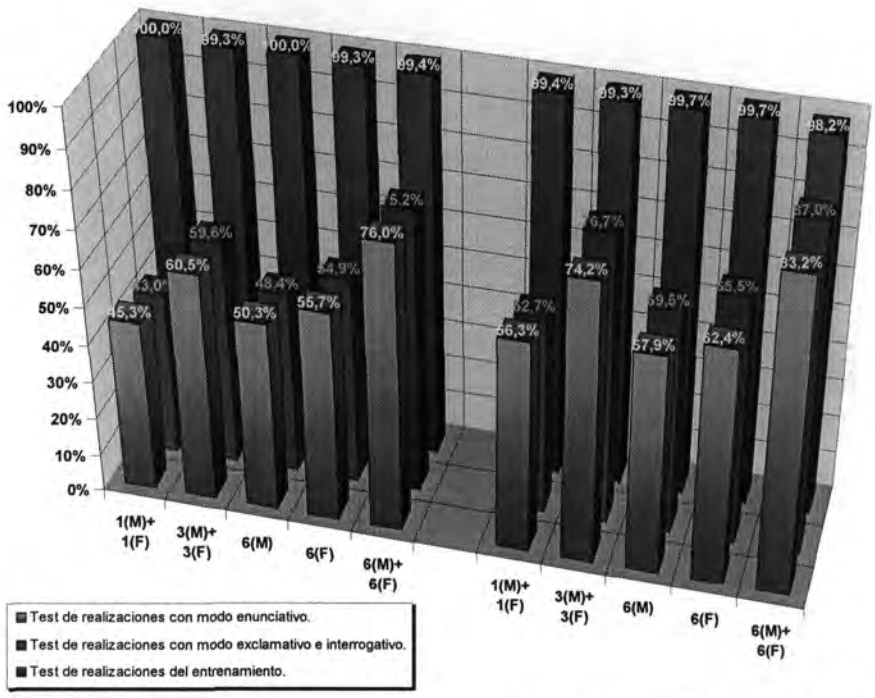

Figura A.89. Tasas de reconocimiento para el experimento PLP-1a (PLP [10 coeficientes]). La parte de la izquierda corresponde a los entrenamientos efectuados solamente con realizaciones con modo enunciativo.

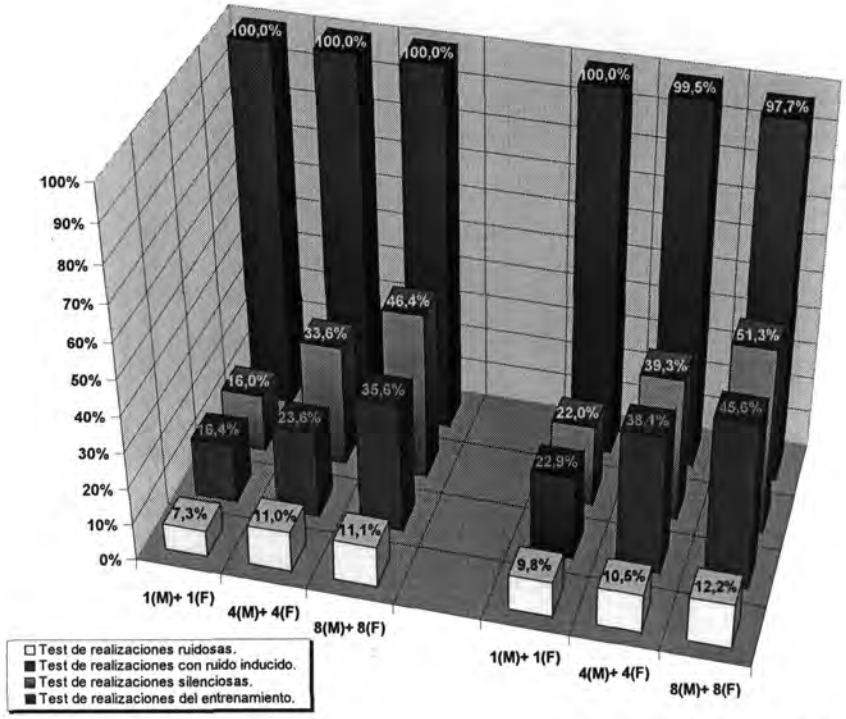

Figura A.90. Tasas de reconocimiento para el experimento PLP-1b (PLP [10 coeficientes]). La parte de la izquierda corresponde a los entrenamientos efectuados solamente con realizaciones silenciosas.

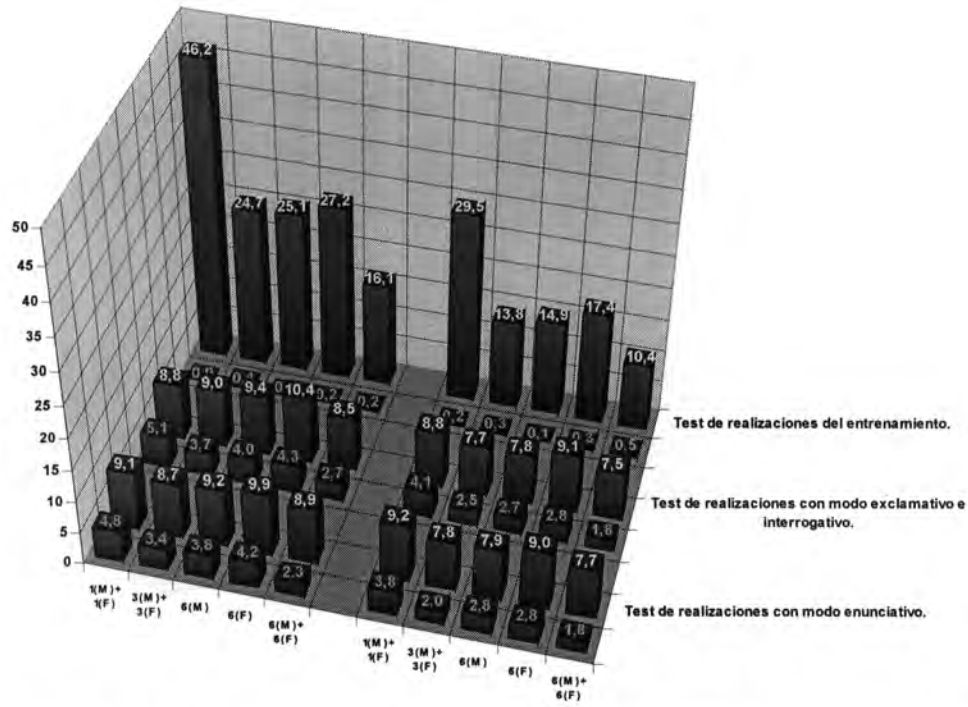

Figura A.91. Tasas de fiabilidad para el experimento PLP-1a (PLP [10 coeficientes]). En verde aparecen las tasas para los casos reconocidos correctamente. En rojo los casos clasificados erróneamente.

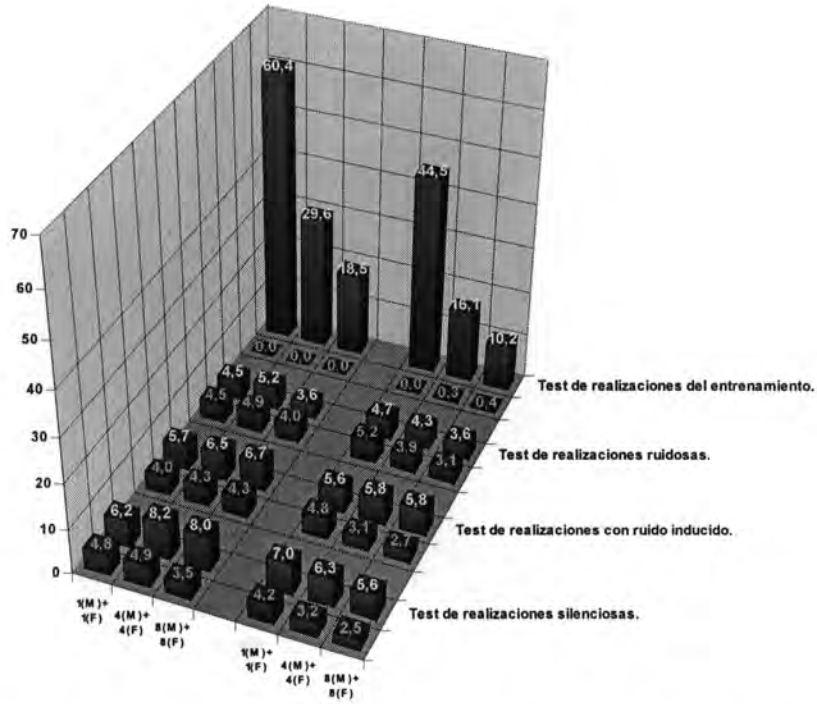

Figura A.92. Tasas de fiabilidad para el experimento PLP-1b (PLP [10 coeficientes]). En verde aparecen las tasas para los casos reconocidos correctamente. En rojo los casos clasificados erróneamente.

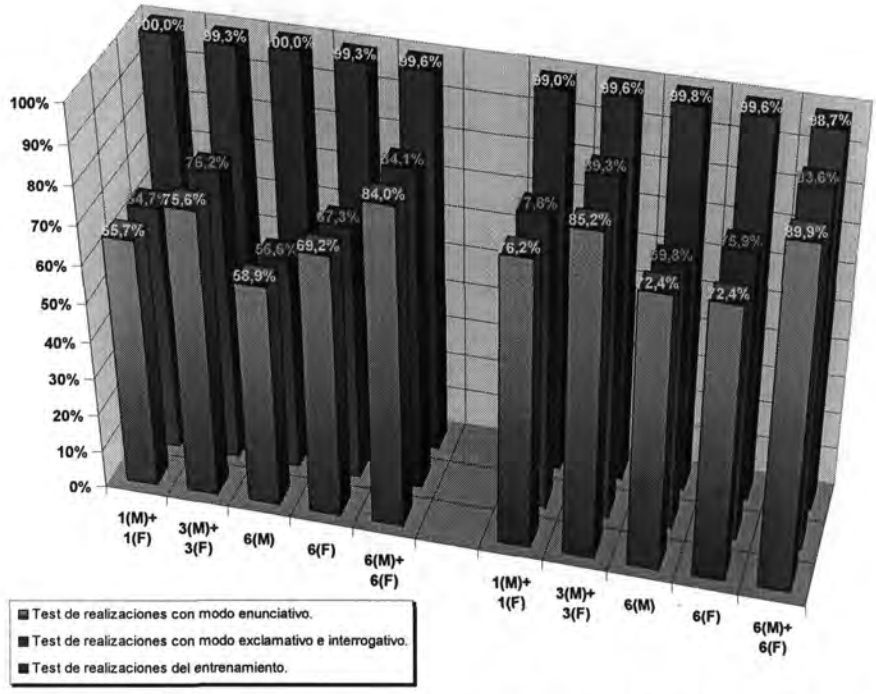

Figura A.93. Tasas de reconocimiento para el experimento PLP-2a (PLP [10 coeficientes + 10 coeficientes delta <5 tramas>]). La parte de la izquierda corresponde a los entrenamientos efectuados solamente con realizaciones con modo enunciativo.

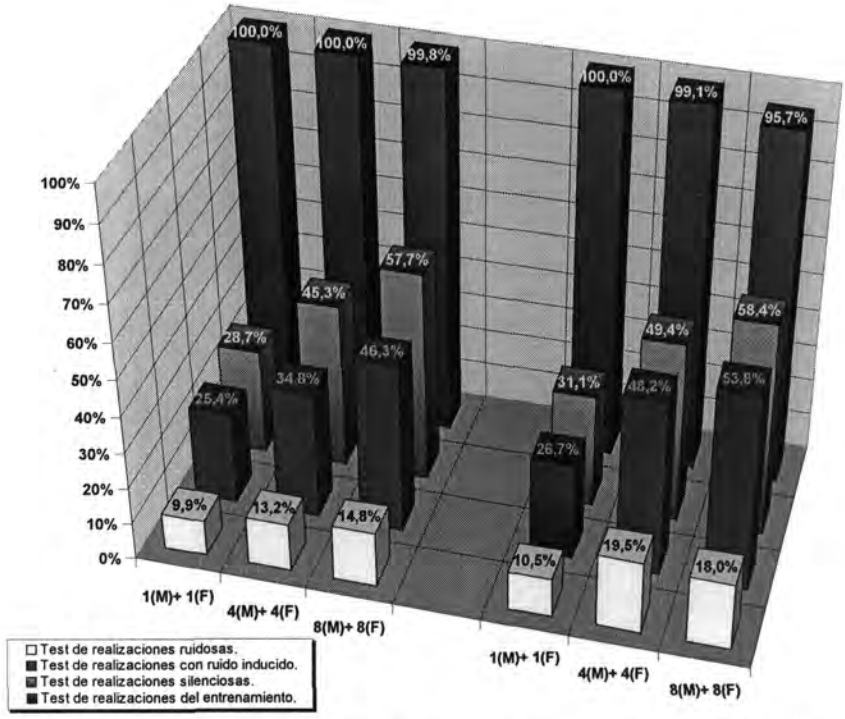

Figura A.94. Tasas de reconocimiento para el experimento PLP-2b (PLP [10 coeficientes + 10 coeficientes delta <5 tramas>]). La parte de la izquierda corresponde a los entrenamientos efectuados solamente con realizaciones silenciosas.

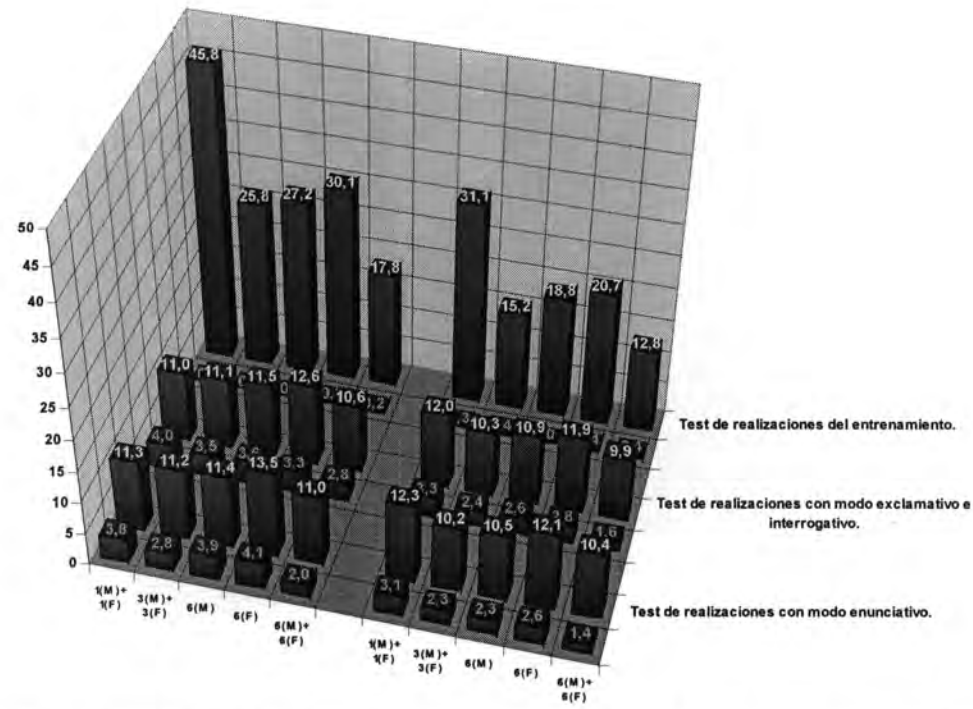

Figura A.95. Tasas de fiabilidad para el experimento PLP-2a (PLP [10 coeficientes + 10 coeficientes delta <5 tramas>]). En verde aparecen las tasas para los casos reconocidos correctamente. En rojo los casos clasificados erróneamente.

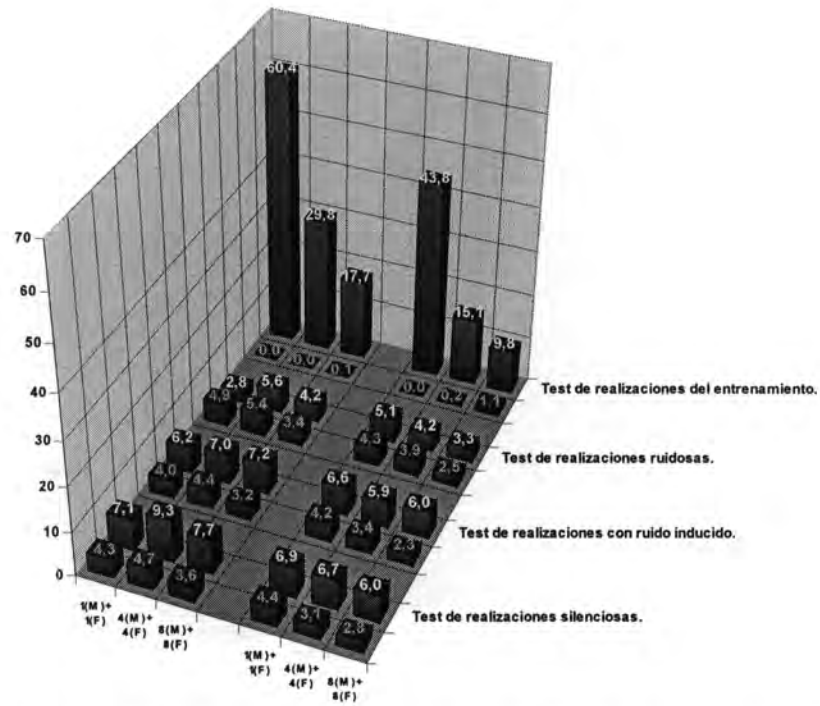

Figura A.96. Tasas de fiabilidad para el experimento PLP-2b (PLP [10 coeficientes + 10 coeficientes delta <5 tramas>]). En verde aparecen las tasas para los casos reconocidos correctamente. En rojo los casos clasificados erróneamente.

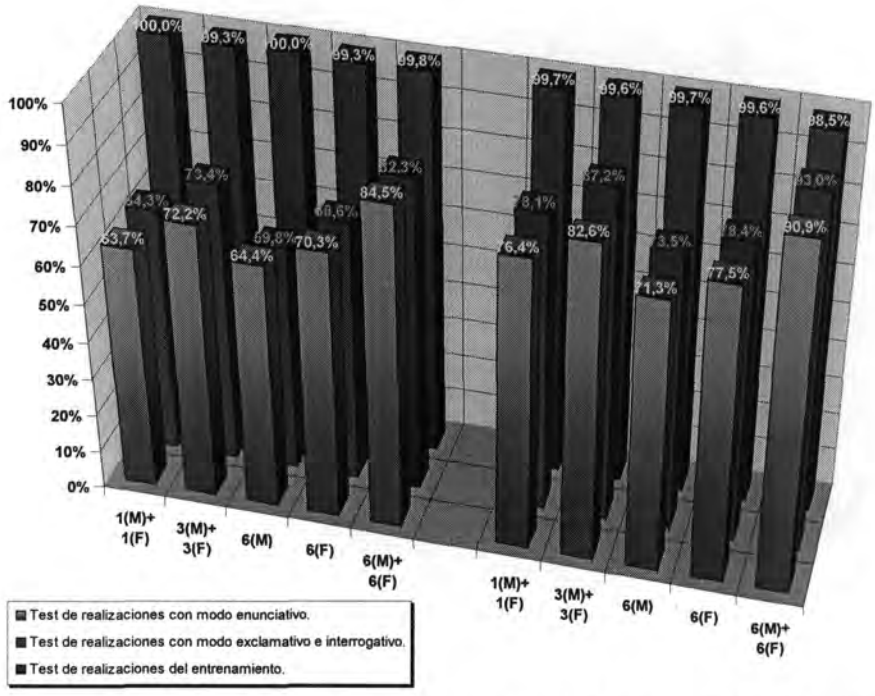

Figura A.97. Tasas de reconocimiento para el experimento PLP-3a (PLP [10 coeficientes + 10 coeficientes delta <5 tramas> + 10 coeficientes delta-delta]). La parte de la izquierda corresponde a los entrenamientos efectuados solamente con realizaciones con modo enunciativo.

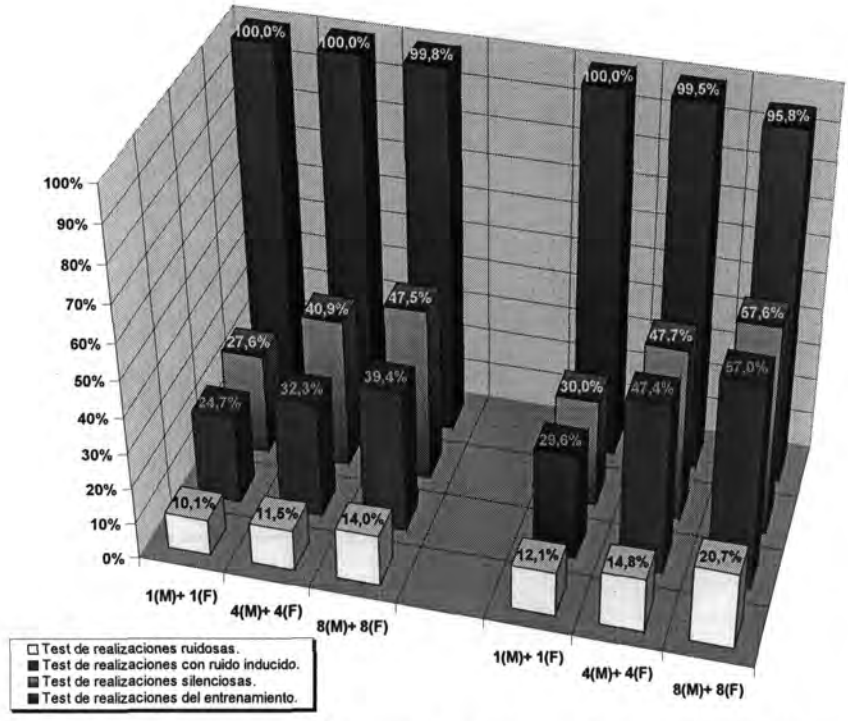

Figura A.98. Tasas de reconocimiento para el experimento PLP-3b (PLP [10 coeficientes + 10 coeficientes delta <5 tramas> + 10 coeficientes delta-delta]). La parte de la izquierda corresponde a los entrenamientos efectuados solamente con realizaciones silenciosas.

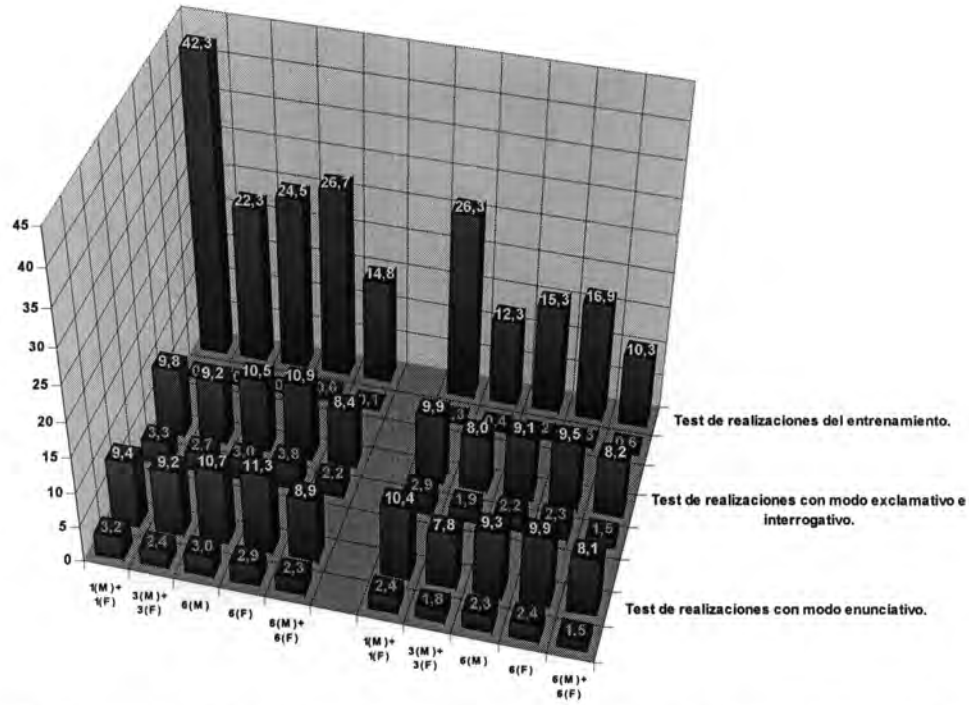

Figura A.99. Tasas de fiabilidad para el experimento PLP-3a (PLP [10 coeficientes + 10 coeficientes delta <5 tramas> + 10 coeficientes delta-delta]). En verde aparecen las tasas para los casos reconocidos correctamente. En rojo los casos clasificados erróneamente.

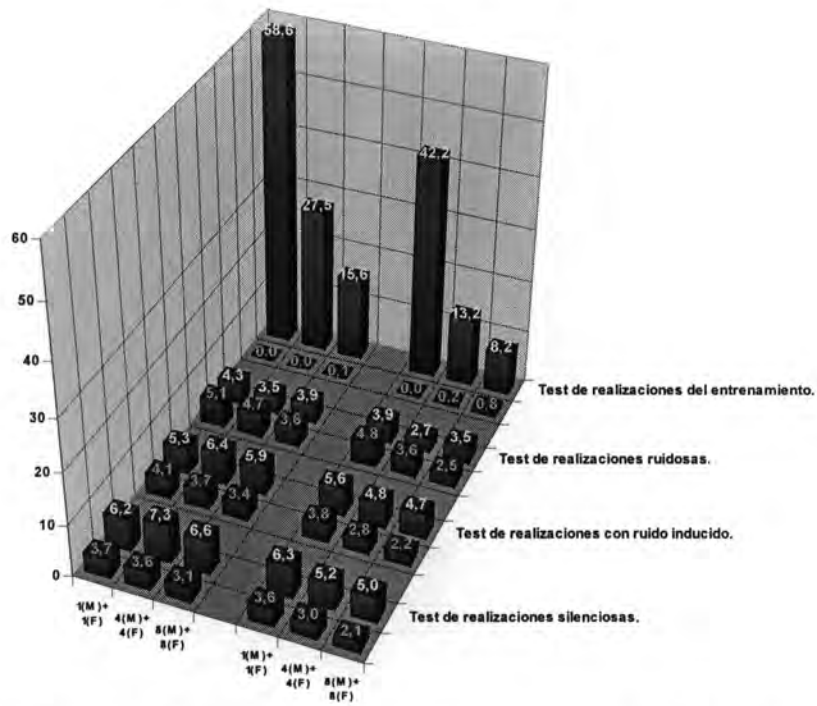

Figura A.100. Tasas de fiabilidad para el experimento PLP-3b (PLP [10 coeficientes + 10 coeficientes delta <5 tramas> + 10 coeficientes delta-delta]). En verde aparecen las tasas para los casos reconocidos correctamente. En rojo los casos clasificados erróneamente.

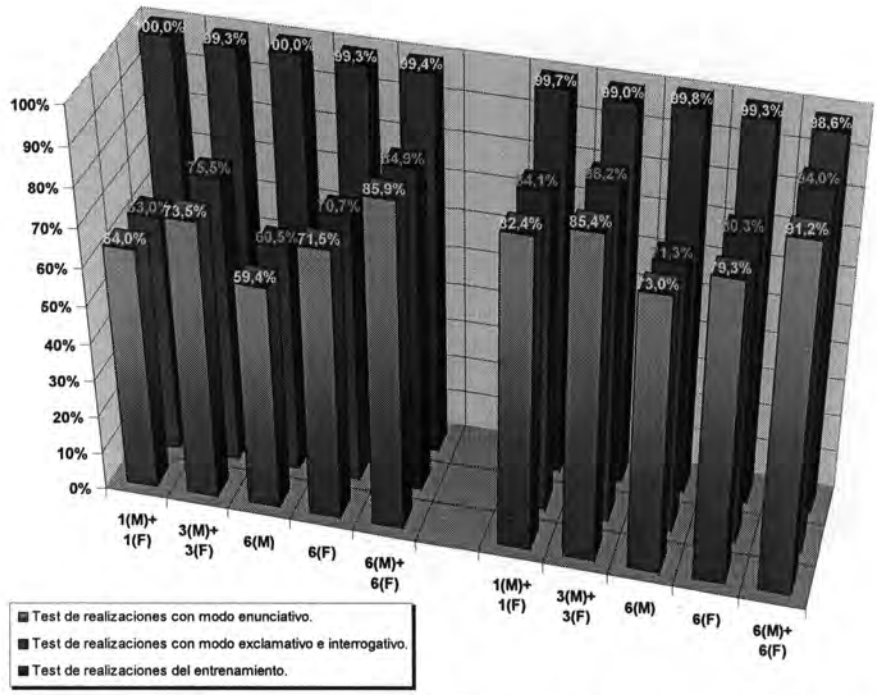

Figura A.101. Tasas de reconocimiento para el experimento PLP-4a (PLP [10 coeficientes + 10 coeficientes delta <5 tramas> + 10 coeficientes delta-delta + energía trama + energía trama diferencia <2 tramas>]). La parte de la izquierda corresponde a los entrenamientos efectuados solamente con realizaciones con modo enunciativo.

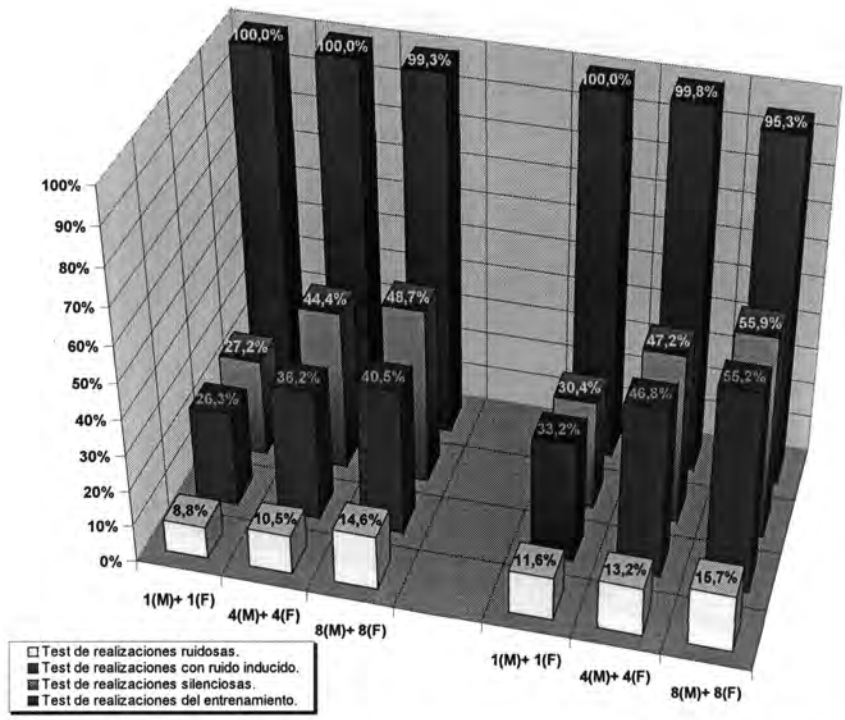

Figura A.102. Tasas de reconocimiento para el experimento PLP-4b (PLP [10 coeficientes + 10 coeficientes delta <5 tramas> + 10 coeficientes delta-delta + energía trama + energía trama diferencia <2 tramas>]). La parte de la izquierda corresponde a los entrenamientos efectuados solamente con realizaciones silenciosas.

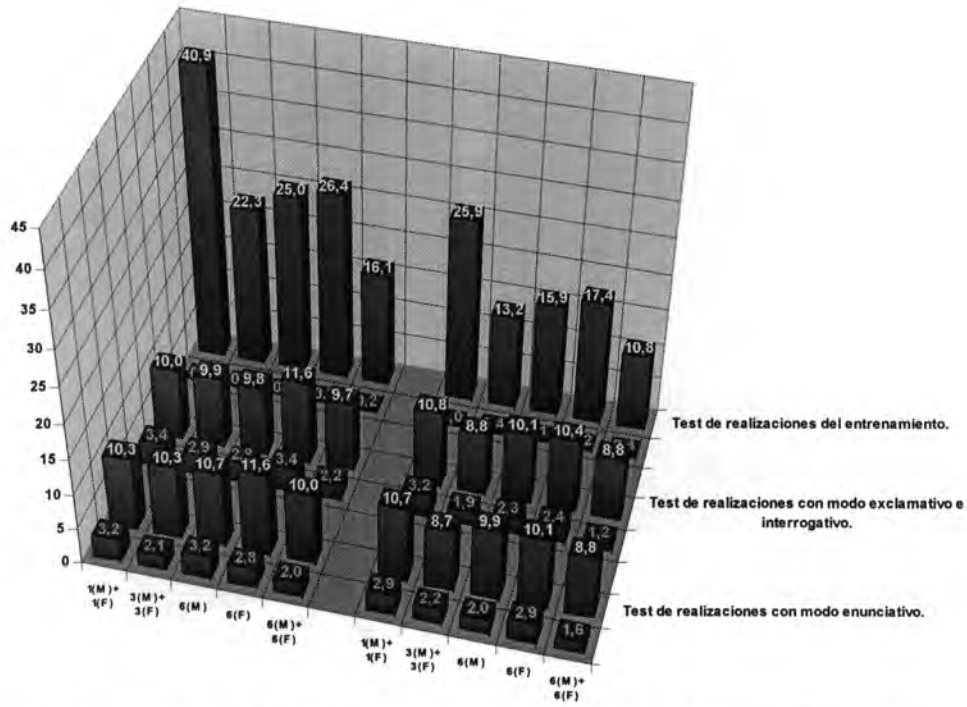

Figura A.103. Tasas de fiabilidad para el experimento PLP-4a (PLP [10 coeficientes + 10 coeficientes delta <5 tramas> + 10 coeficientes delta-delta + energía trama + energía trama diferencia <2 tramas>]). En verde aparecen las tasas para los casos reconocidos correctamente. En rojo los casos clasificados erróneamente.

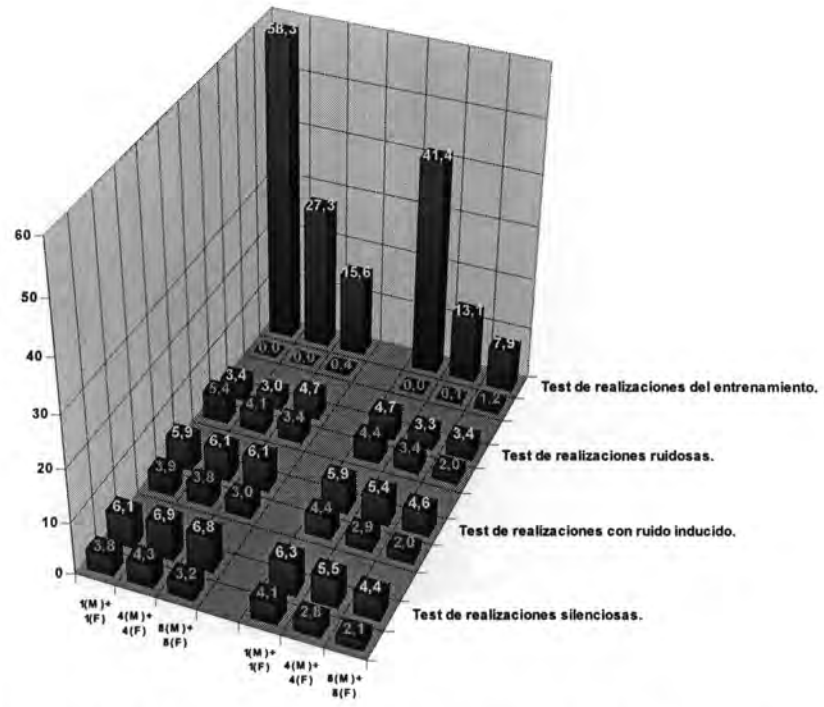

Figura A.104. Tasas de fiabilidad para el experimento PLP-4b (PLP [10 coeficientes + 10 coeficientes delta <5 tramas> + 10 coeficientes delta-delta + energía trama + energía trama diferencia <2 tramas>]). En verde aparecen las tasas para los casos reconocidos correctamente. En rojo los casos clasificados erróneamente.

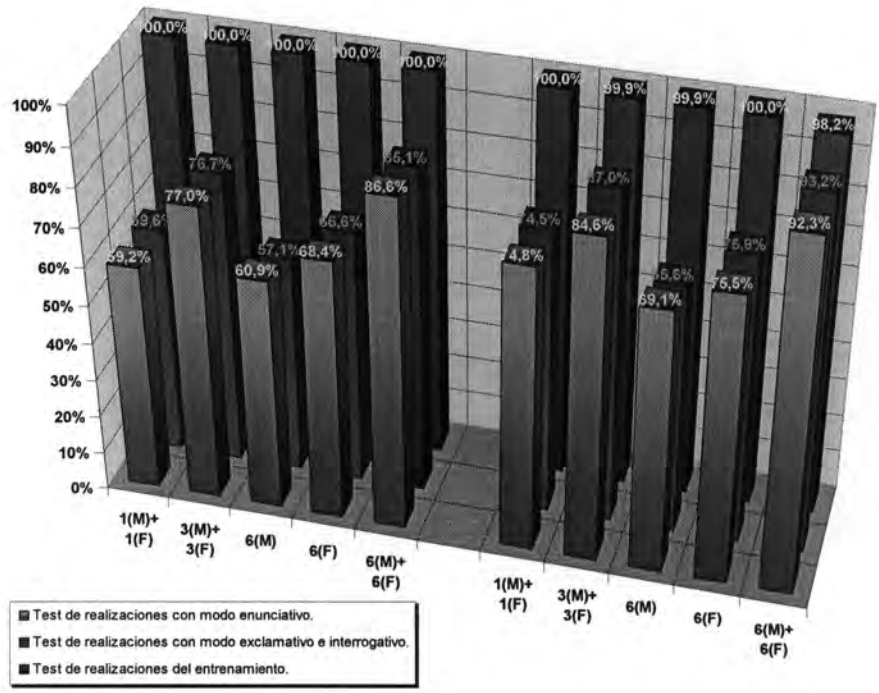

Figura A.105. Tasas de reconocimiento para el experimento PLP-5a (PLP sin normalización de energía en LPC-Cepstrum [10 coeficientes + 10 coeficientes delta <5 tramas>]). La parte de la izquierda corresponde a los entrenamientos efectuados solamente con realizaciones con modo enunciativo.

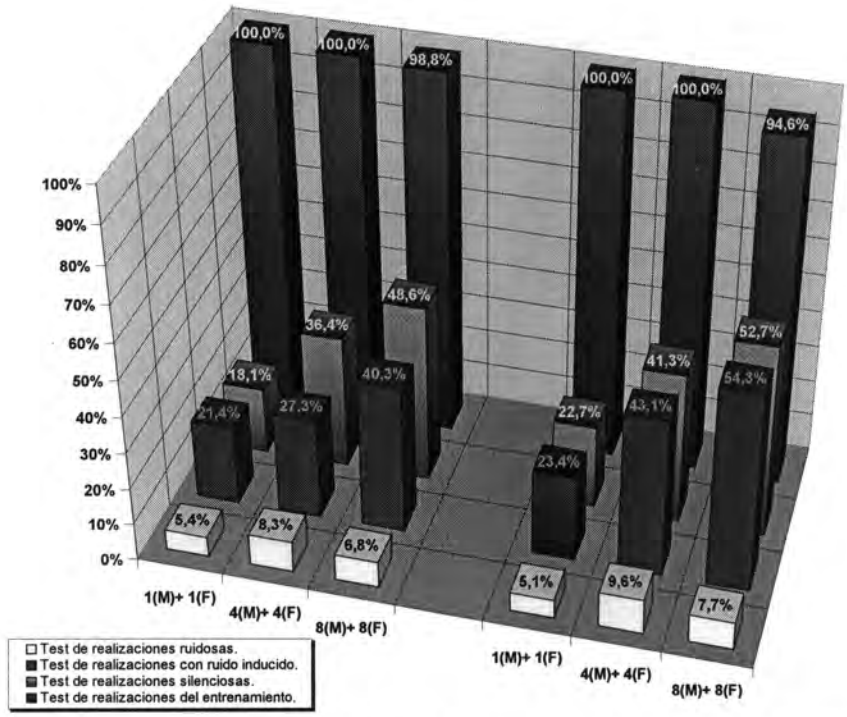

Figura A.106. Tasas de reconocimiento para el experimento PLP-5b (PLP sin normalización de energía en LPC-Cepstrum [10 coeficientes + 10 coeficientes delta <5 tramas>]). La parte de la izquierda corresponde a los entrenamientos efectuados solamente con realizaciones silenciosas.

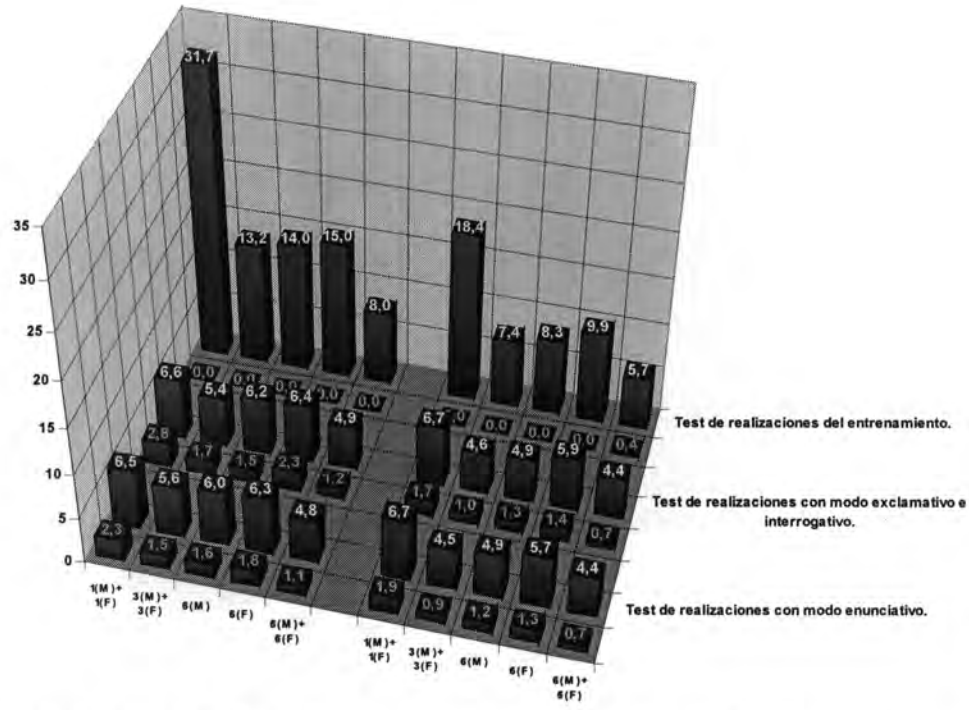

Figura A.107. Tasas de fiabilidad para el experimento PLP-5a (PLP sin normalización de energía en LPC-Cepstrum [10 coeficientes + 10 coeficientes delta <5 tramas>]). En verde aparecen las tasas para los casos reconocidos correctamente. En rojo los casos clasificados erróneamente.

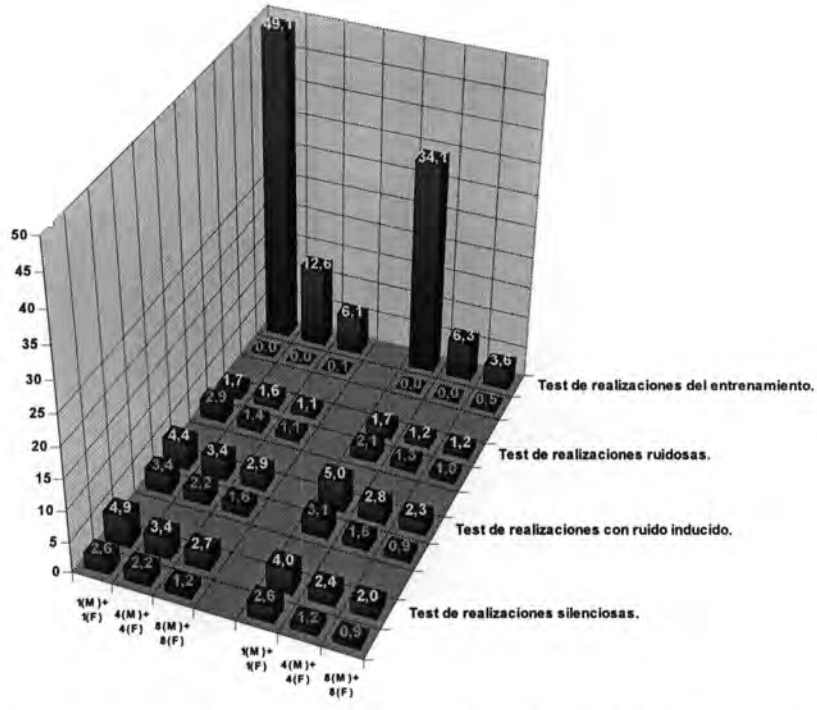

Figura A.108. Tasas de fiabilidad para el experimento PLP-5b (PLP sin normalización de energía en LPC-Cepstrum [10 coeficientes + 10 coeficientes delta <5 tramas>]). En verde aparecen las tasas para los casos reconocidos correctamente. En rojo los casos clasificados erróneamente.

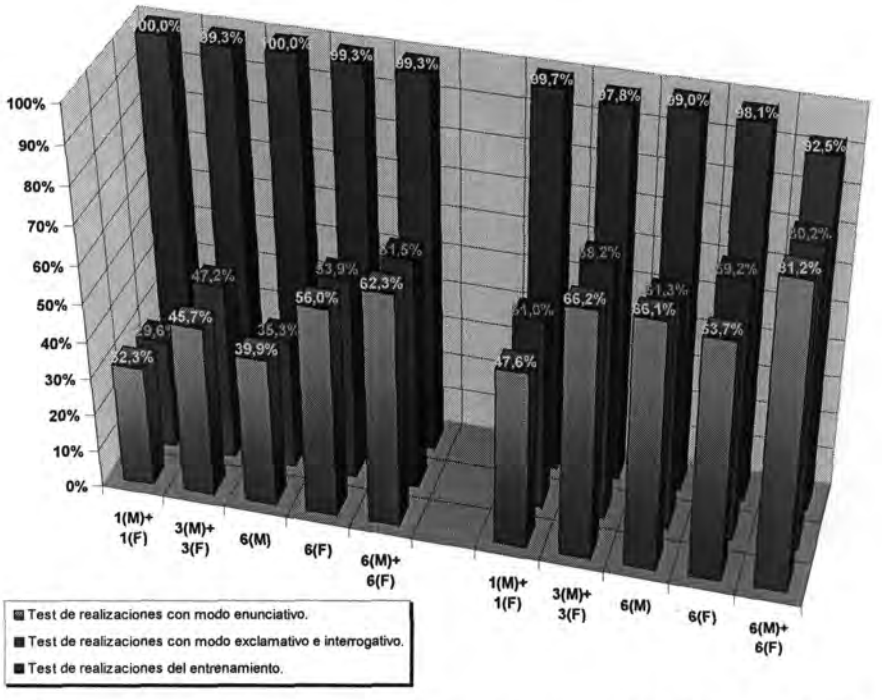

Figura A.109. Tasas de reconocimiento para el experimento PLP-6a (PLP sin LPC-Cepstrum [10 coeficientes + 10 coeficientes delta <5 tramas>]). La parte de la izquierda corresponde a los entrenamientos efectuados solamente con realizaciones con modo enunciativo.

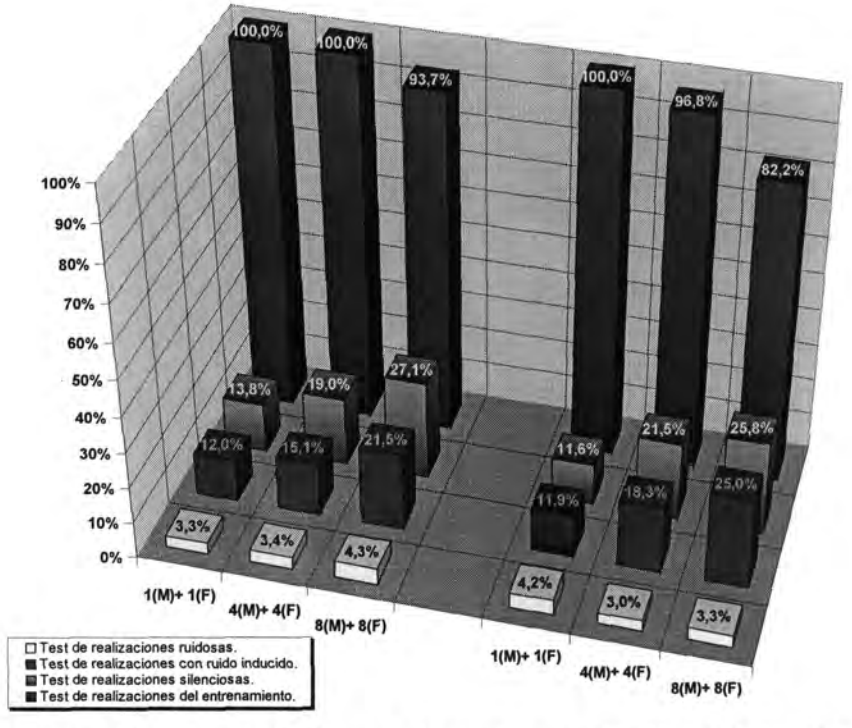

Figura A.110. Tasas de reconocimiento para el experimento PLP-6b (PLP sin LPC-Cepstrum [10 coeficientes + 10 coeficientes delta <5 tramas>]). La parte de la izquierda corresponde a los entrenamientos efectuados solamente con realizaciones silenciosas.

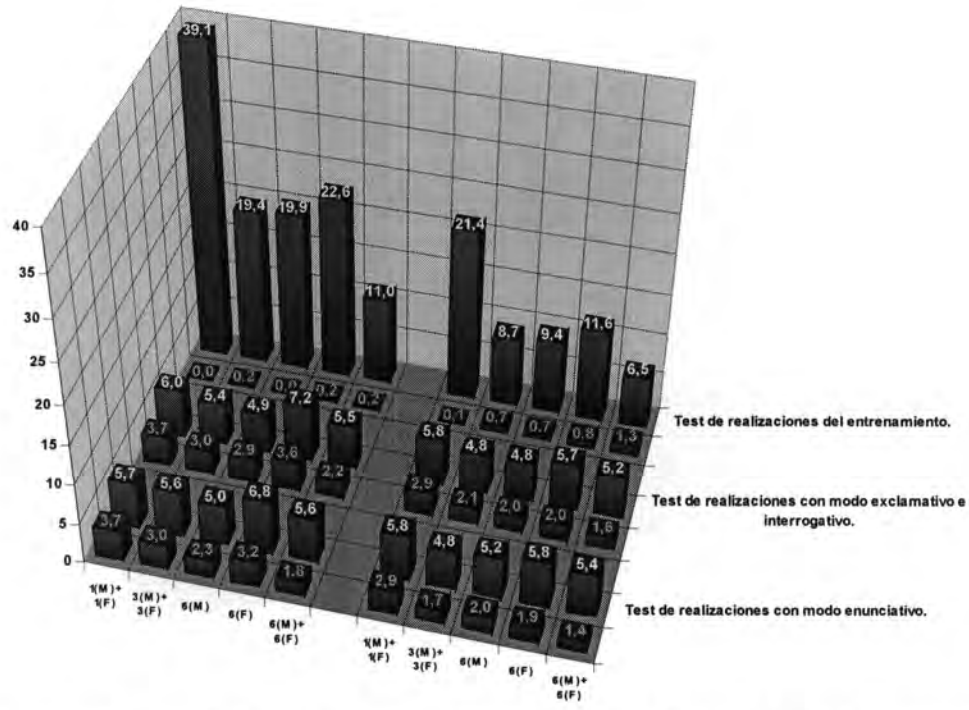

Figura A.111. Tasas de fiabilidad para el experimento PLP-6a (PLP sin LPC-Cepstrum [10 coeficientes + 10 coeficientes delta <5 tramas>]). En verde aparecen las tasas para los casos reconocidos correctamente. En rojo los casos clasificados erróneamente.

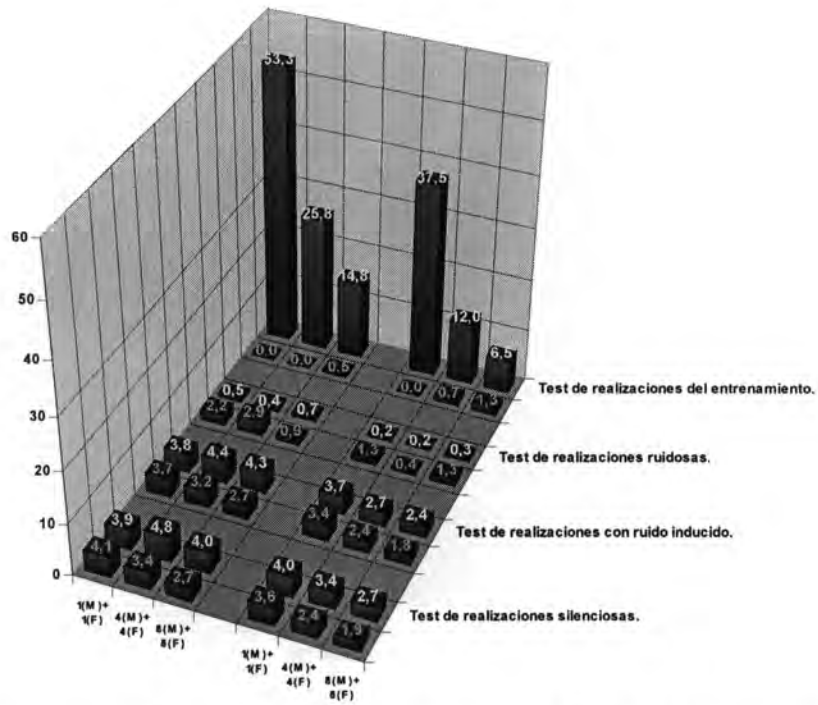

Figura A.112. Tasas de fiabilidad para el experimento PLP-6b (PLP sin LPC-Cepstrum [10 coeficientes + 10 coeficientes delta <5 tramas>]). En verde aparecen las tasas para los casos reconocidos correctamente. En rojo los casos clasificados erróneamente.

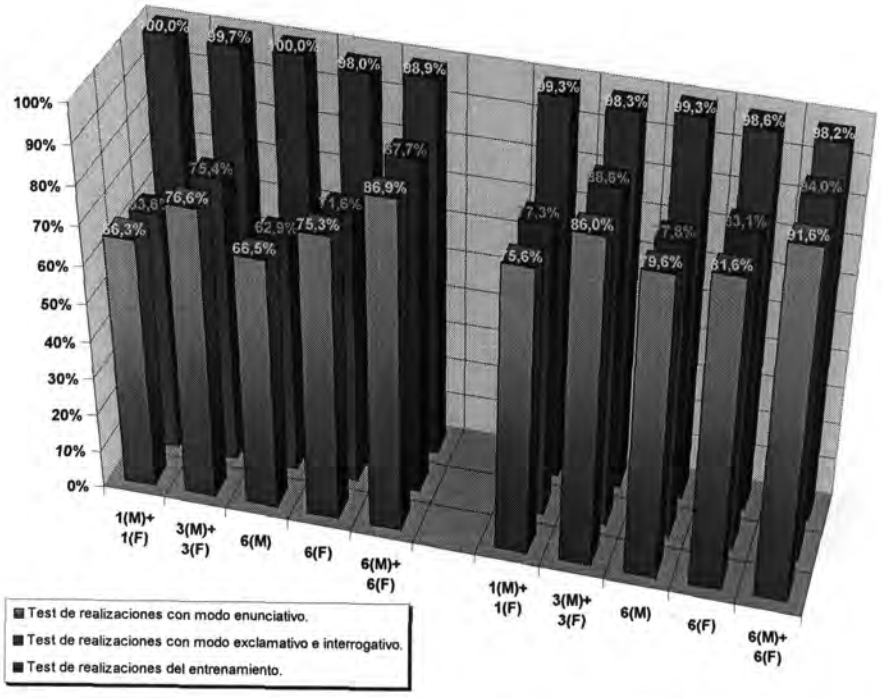

Figura A.113. Tasas de reconocimiento para el experimento PLP-7a (PLP sin tratamiento perceptual [10 coeficientes + 10 coeficientes delta <5 tramas>]). La parte de la izquierda corresponde a los entrenamientos efectuados solamente con realizaciones con modo enunciativo.

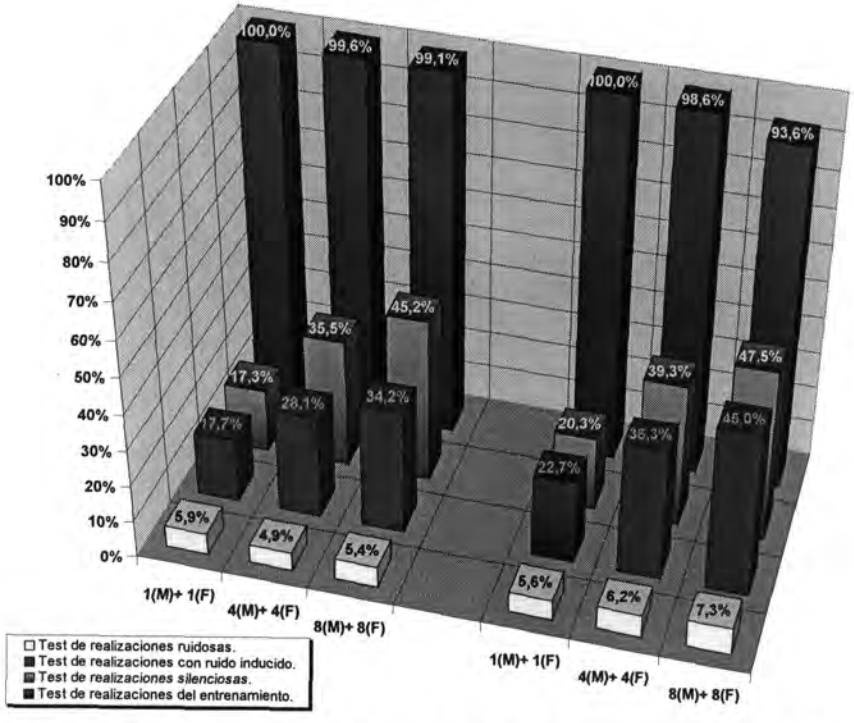

Figura A.114. Tasas de reconocimiento para el experimento PLP-7b (PLP sin tratamiento perceptual [10 coeficientes + 10 coeficientes delta <5 tramas>]). La parte de la izquierda corresponde a los entrenamientos efectuados solamente con realizaciones silenciosas.

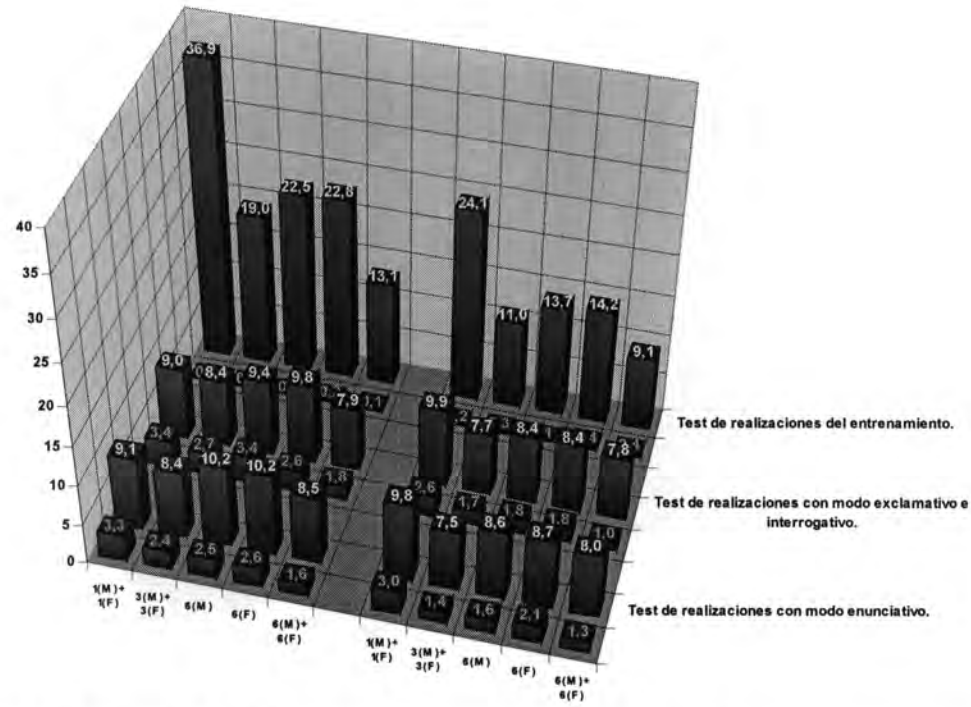

Figura A.115. Tasas de fiabilidad para el experimento PLP-7a (PLP sin tratamiento perceptual [10 coeficientes + 10 coeficientes delta <5 tramas>]). En verde aparecen las tasas para los casos reconocidos correctamente. En rojo los casos clasificados erróneamente.

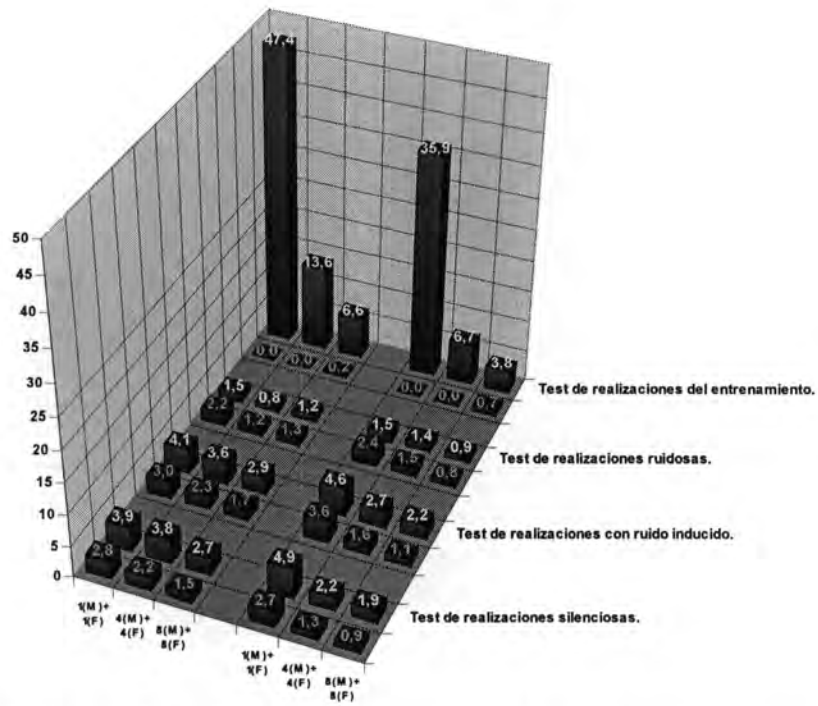

Figura A.116. Tasas de fiabilidad para el experimento PLP-7b (PLP sin tratamiento perceptual [10 coeficientes + 10 coeficientes delta <5 tramas>]). En verde aparecen las tasas para los casos reconocidos correctamente. En rojo los casos clasificados erróneamente.

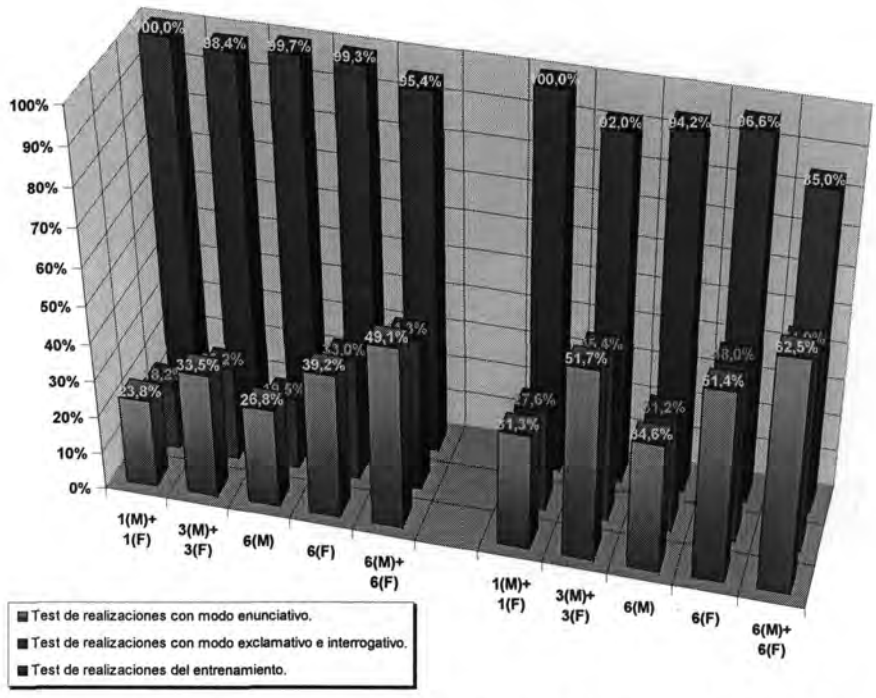

Figura A.117. Tasas de reconocimiento para el experimento PLP-8a (P-Cepstrum [10 coeficientes]). La parte de la izquierda corresponde a los entrenamientos efectuados solamente con realizaciones con modo enunciativo.

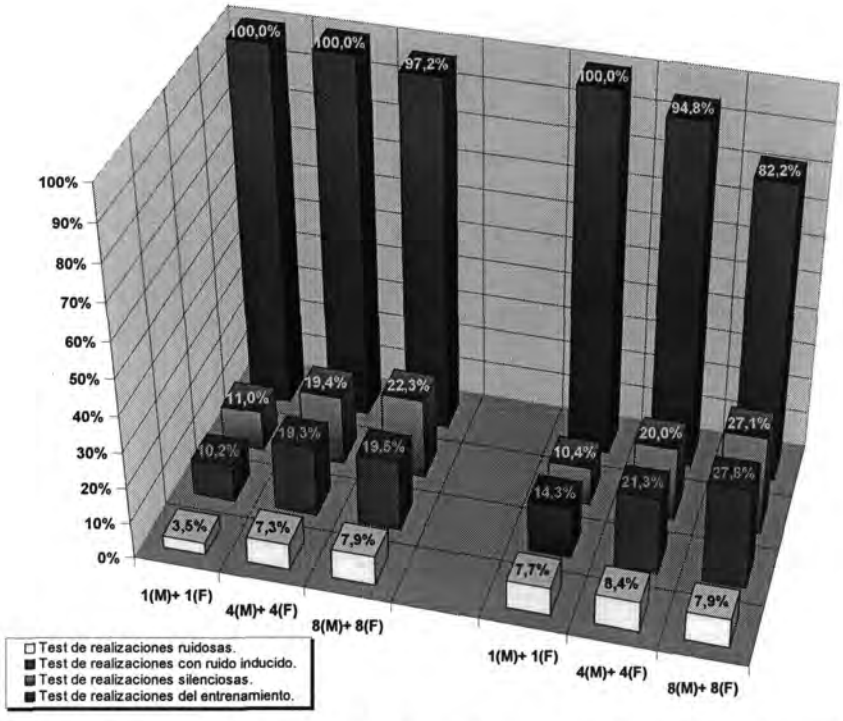

Figura A.118. Tasas de reconocimiento para el experimento PLP-8b (P-Cepstrum [10 coeficientes]). La parte de la izquierda corresponde a los entrenamientos efectuados solamente con realizaciones silenciosas.

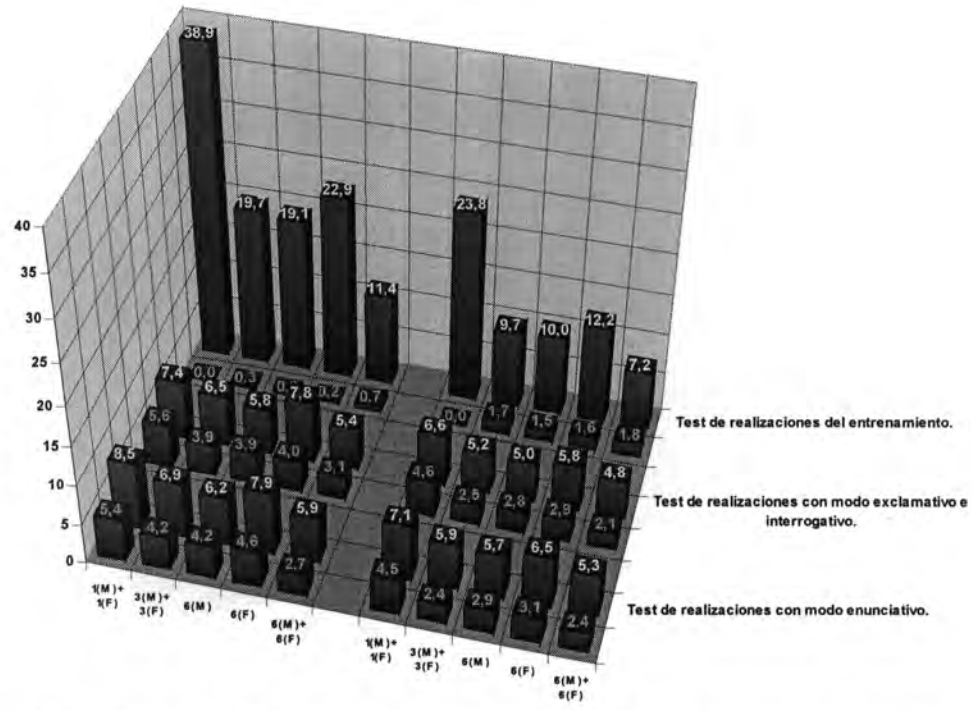

Figura A.119. Tasas de fiabilidad para el experimento PLP-8a (P-Cepstrum [10 coeficientes]). En verde aparecen las tasas para los casos reconocidos correctamente. En rojo los casos clasificados erróneamente.

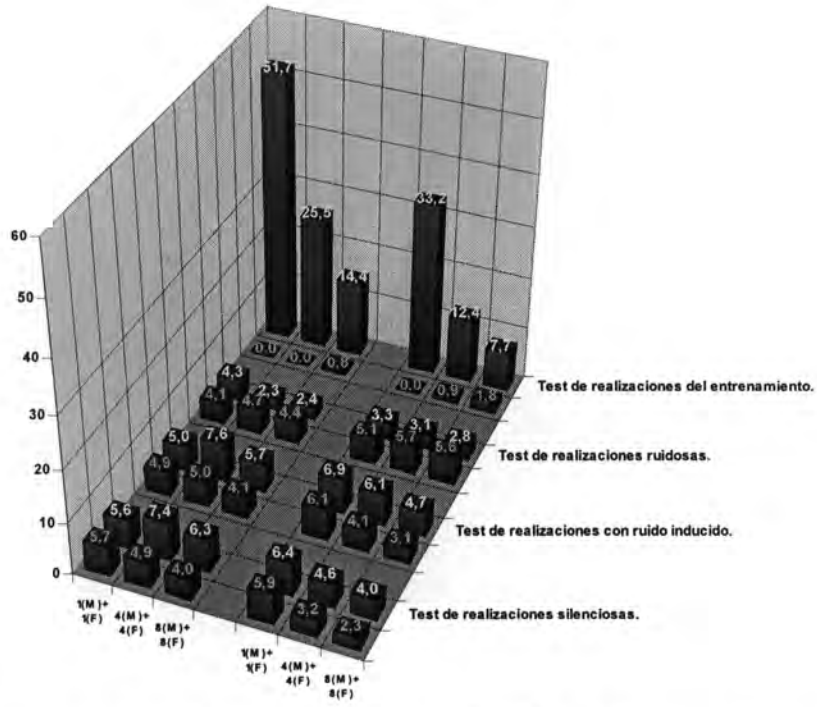

Figura A.120. Tasas de fiabilidad para el experimento PLP-8b (P-Cepstrum [10 coeficientes]). En verde aparecen las tasas para los casos reconocidos correctamente. En rojo los casos clasificados erróneamente.

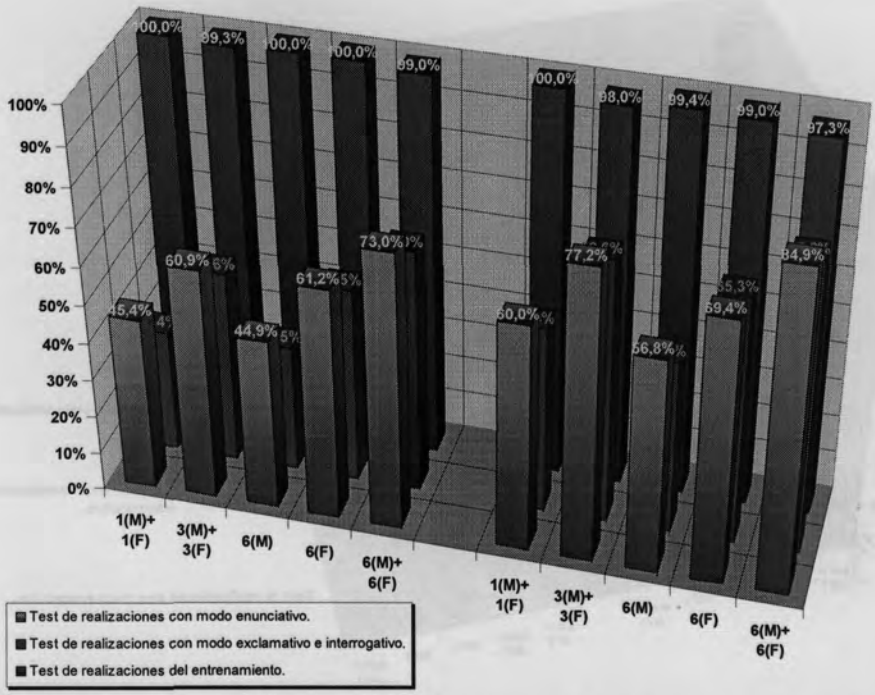

Figura A.121. Tasas de reconocimiento para el experimento PLP-9a (P-Cepstrum [10 coeficientes + 10 coeficientes delta <5 tramas>]). La parte de la izquierda corresponde a los entrenamientos efectuados solamente con realizaciones con modo enunciativo.

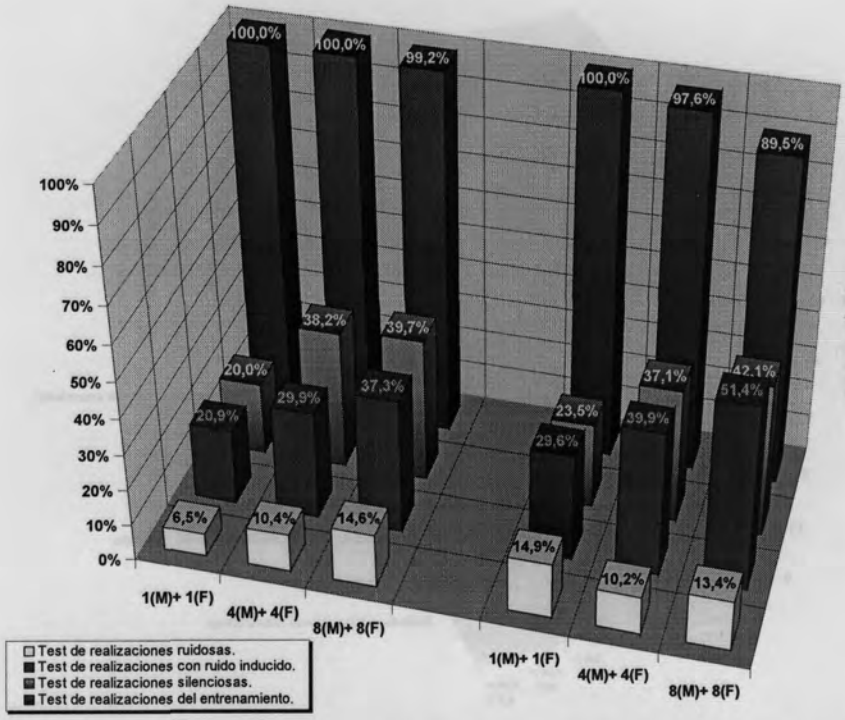

Figura A.122. Tasas de reconocimiento para el experimento PLP-9b (P-Cepstrum [10 coeficientes + 10 coeficientes delta <5 tramas>]). La parte de la izquierda corresponde a los entrenamientos efectuados solamente con realizaciones silenciosas.

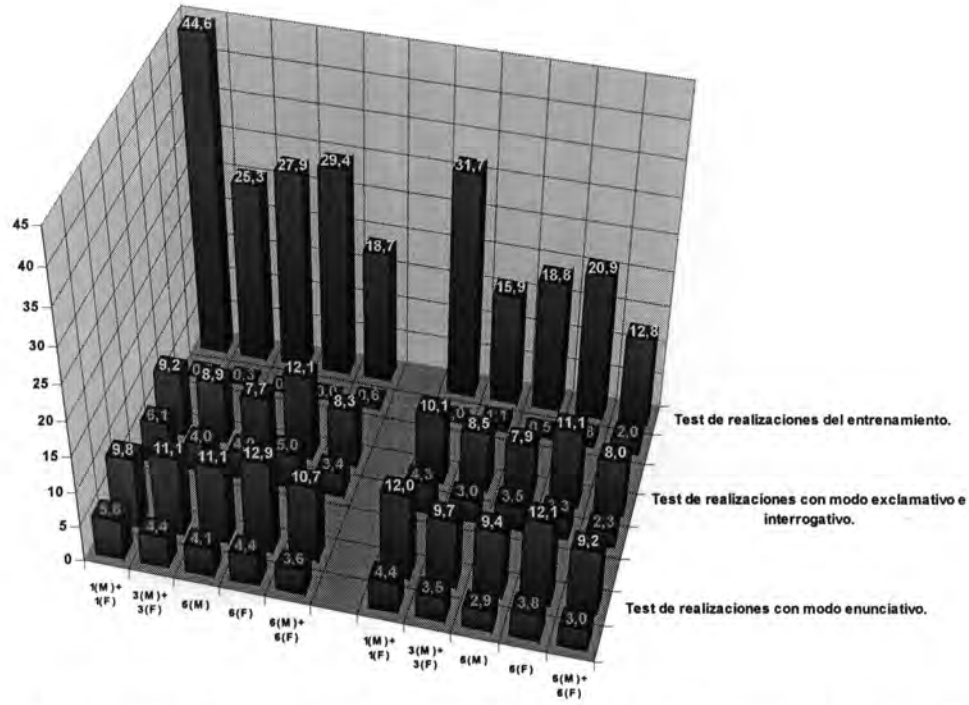

Figura A.123. Tasas de fiabilidad para el experimento PLP-9a (P-Cepstrum [10 coeficientes + 10 coeficientes delta <5 tramas>]). En verde aparecen las tasas para los casos reconocidos correctamente. En rojo los casos clasificados erróneamente.

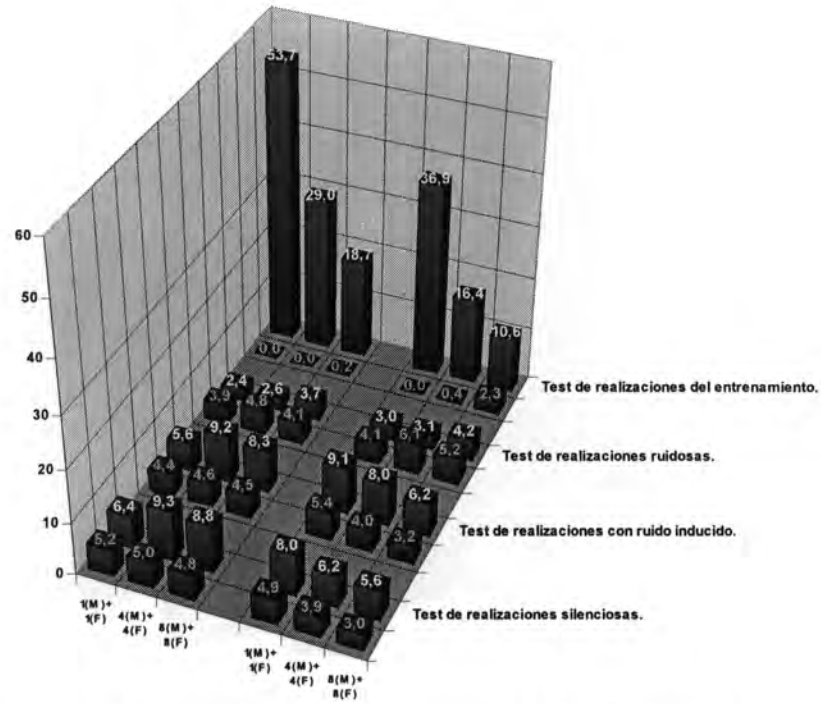

Figura A.124. Tasas de fiabilidad para el experimento PLP-9b (P-Cepstrum [10 coeficientes + 10 coeficientes delta <5 tramas>]). En verde aparecen las tasas para los casos reconocidos correctamente. En rojo los casos clasificados erróneamente.

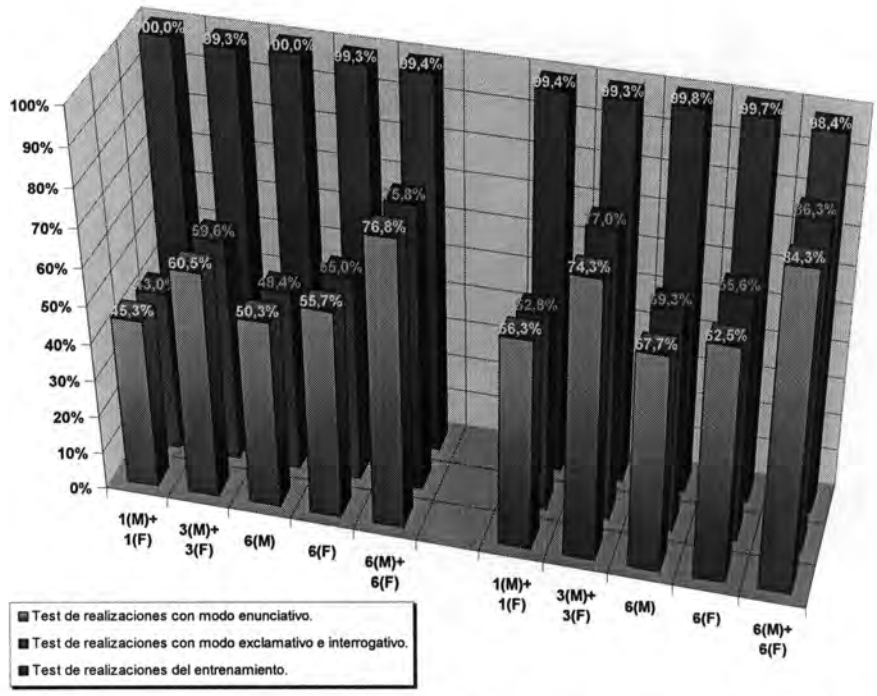

Figura A.125. Tasas de reconocimiento para el experimento PLP-10a (RASTA-PLP [10 coeficientes]). La parte de la izquierda corresponde a los entrenamientos efectuados solamente con realizaciones con modo enunciativo.

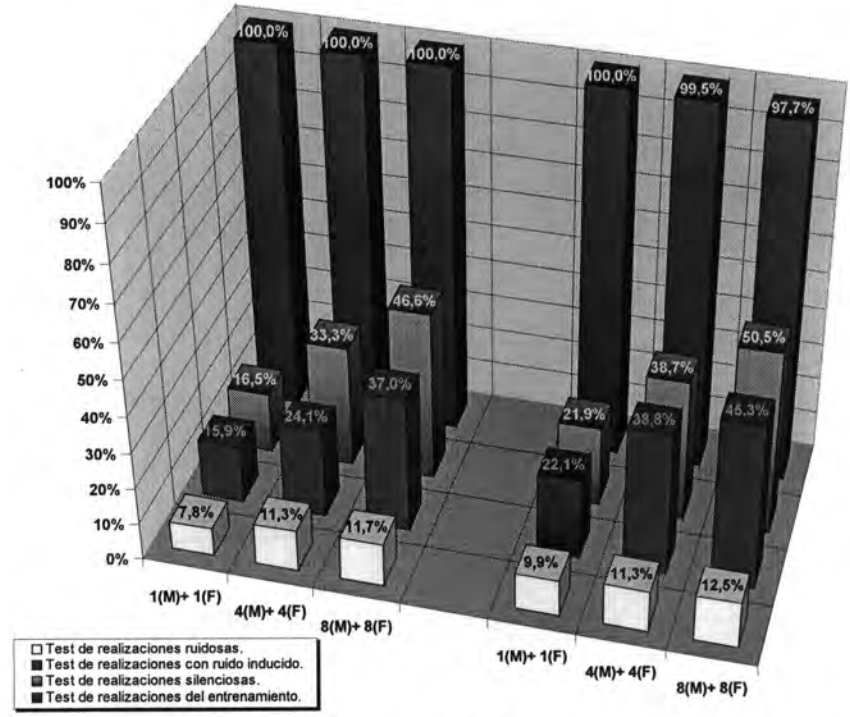

Figura A.126. Tasas de reconocimiento para el experimento PLP-10b (RASTA-PLP [10 coeficientes]). La parte de la izquierda corresponde a los entrenamientos efectuados solamente con realizaciones silenciosas.

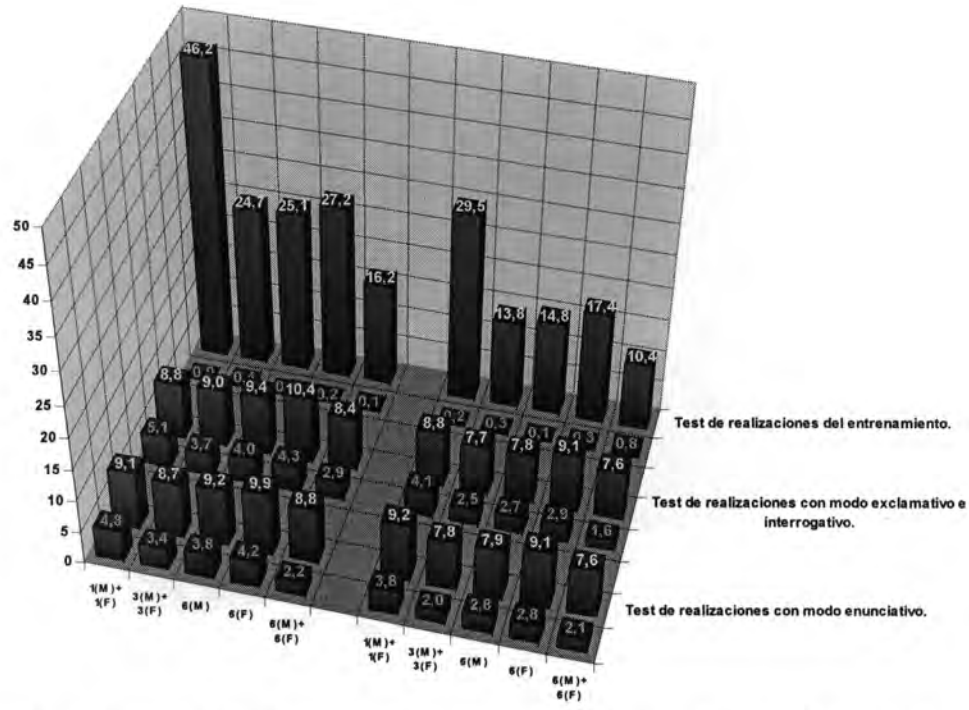

Figura A.127. Tasas de fiabilidad para el experimento PLP-10a (RASTA-PLP [10 coeficientes]). En verde aparecen las tasas para los casos reconocidos correctamente. En rojo los casos clasificados erróneamente.

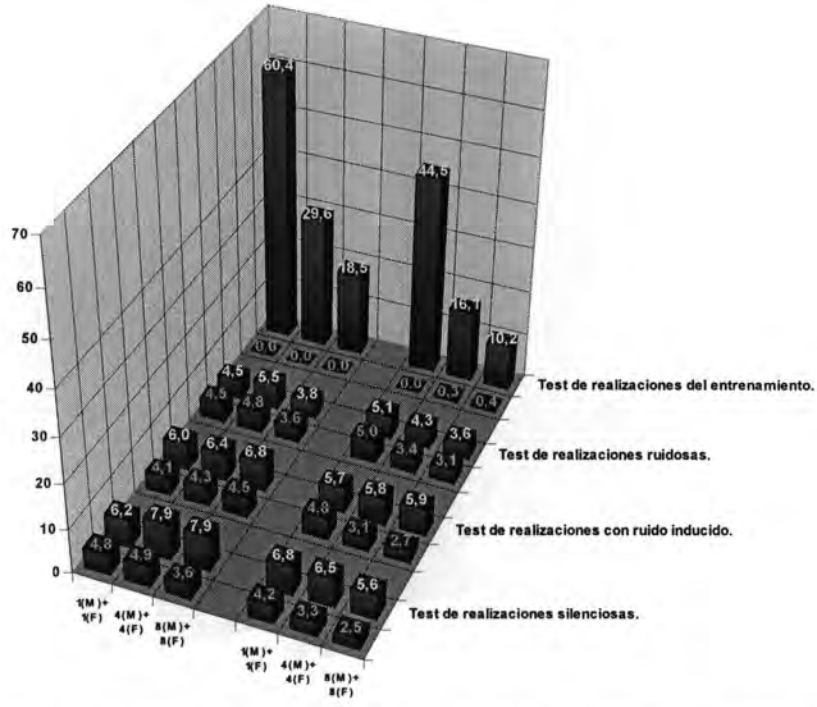

Figura A.128. Tasas de fiabilidad para el experimento PLP-10b (RASTA-PLP [10 coeficientes]). En verde aparecen las tasas para los casos reconocidos correctamente. En rojo los casos clasificados erróneamente.

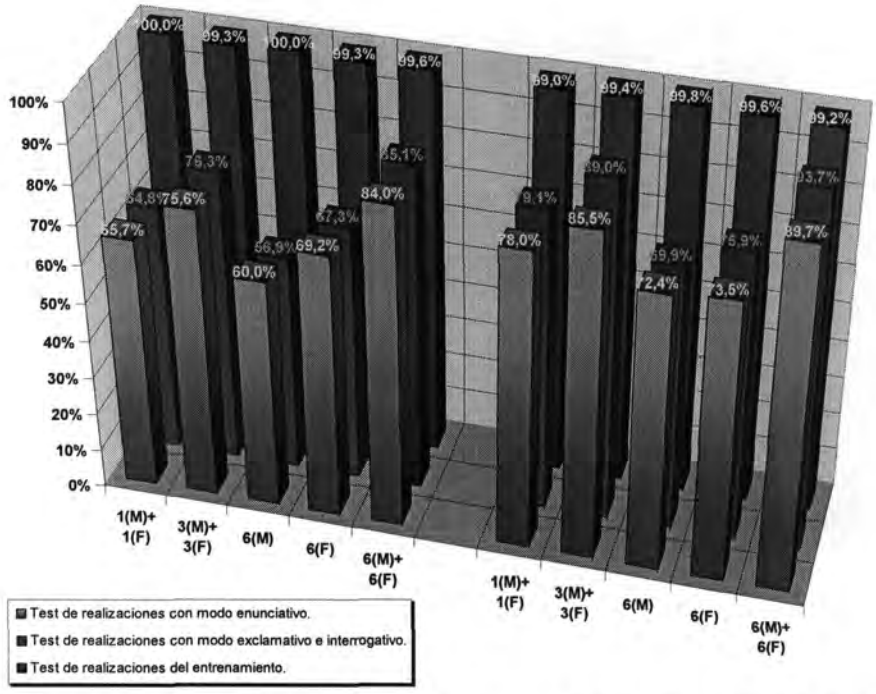

Figura A.129. Tasas de reconocimiento para el experimento PLP-11a (RASTA-PLP [10 coeficientes + 10 coeficientes delta <5 tramas>]). La parte de la izquierda corresponde a los entrenamientos efectuados solamente con realizaciones con modo enunciativo.

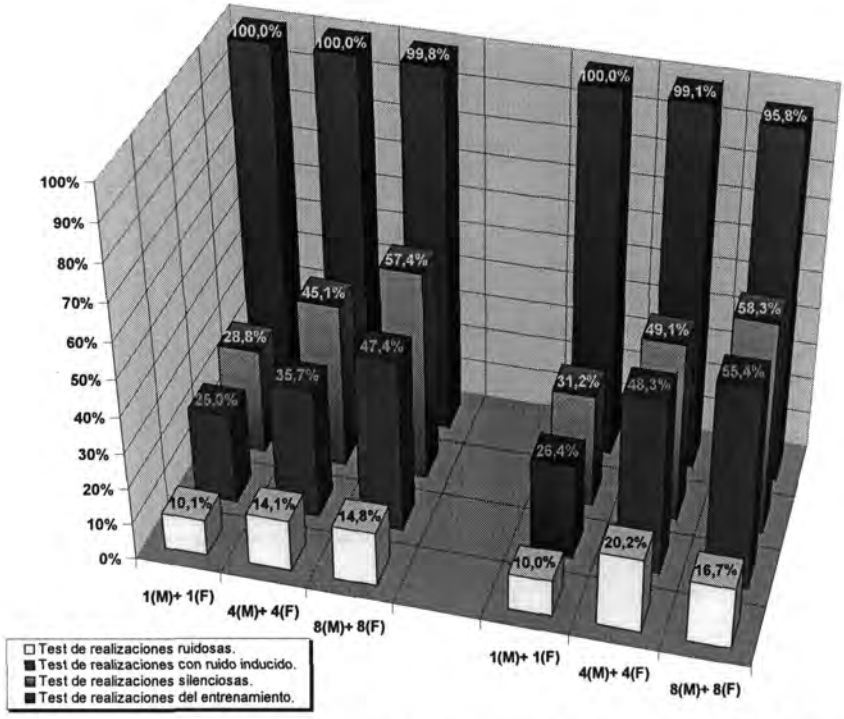

Figura A.130. Tasas de reconocimiento para el experimento PLP-11b (RASTA-PLP [10 coeficientes + 10 coeficientes delta <5 tramas>]). La parte de la izquierda corresponde a los entrenamientos efectuados solamente con realizaciones silenciosas.

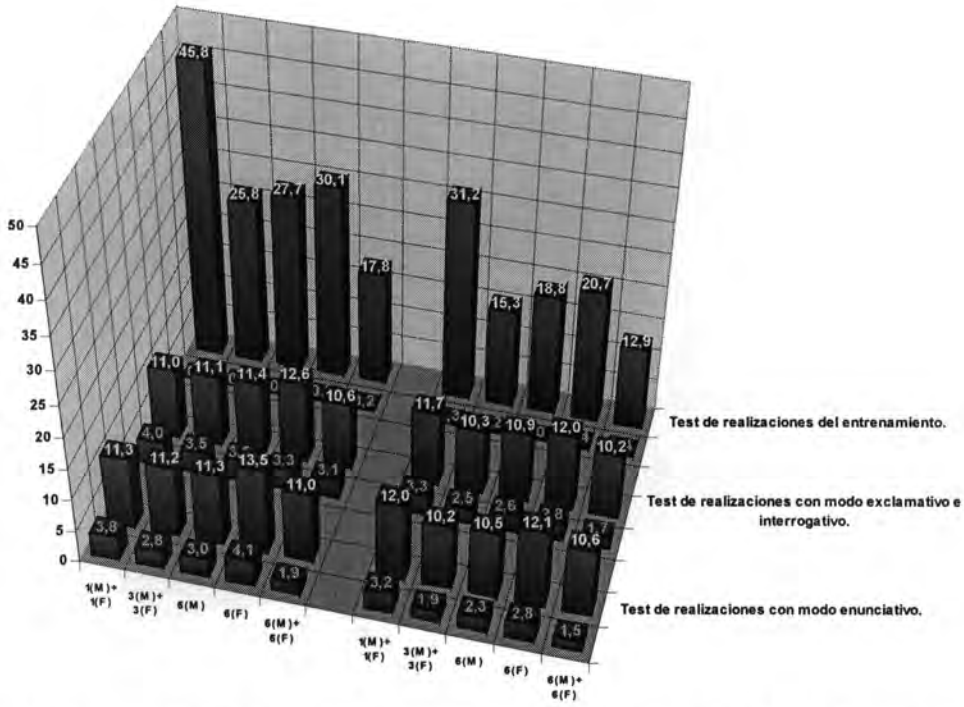

Figura A.131. Tasas de fiabilidad para el experimento PLP-11a (RASTA-PLP [10 coeficientes + 10 coeficientes delta <5 tramas>]). En verde aparecen las tasas para los casos reconocidos correctamente. En rojo los casos clasificados erróneamente.

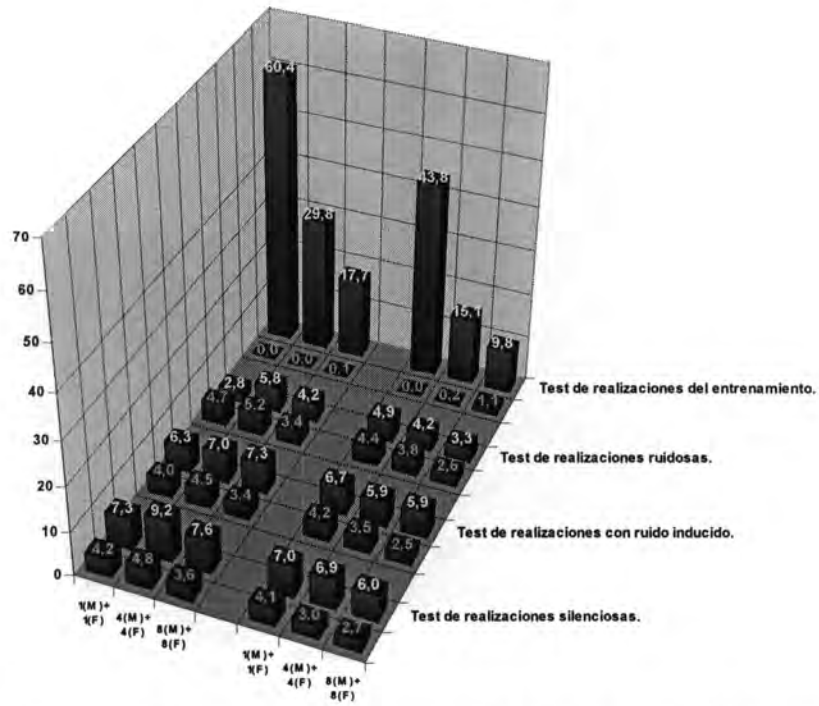

Figura A.132. Tasas de fiabilidad para el experimento PLP-11b (RASTA-PLP [10 coeficientes + 10 coeficientes delta <5 tramas>]). En verde aparecen las tasas para los casos reconocidos correctamente. En rojo los casos clasificados erróneamente.

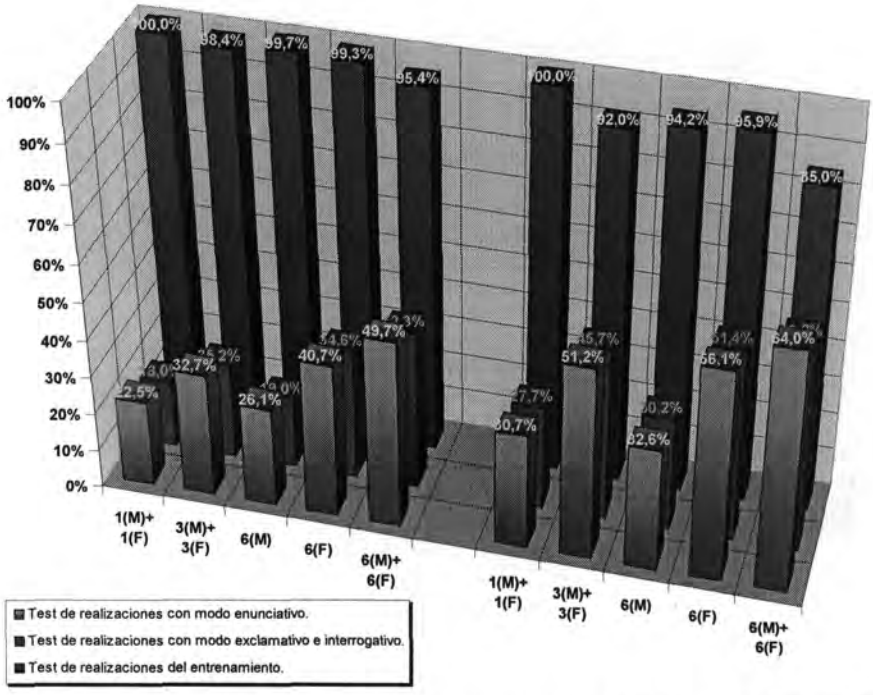

Figura A.133. Tasas de reconocimiento para el experimento PLP-12a (RASTA-P-Cepstrum [10 coeficientes]). La parte de la izquierda corresponde a los entrenamientos efectuados solamente con realizaciones con modo enunciativo.

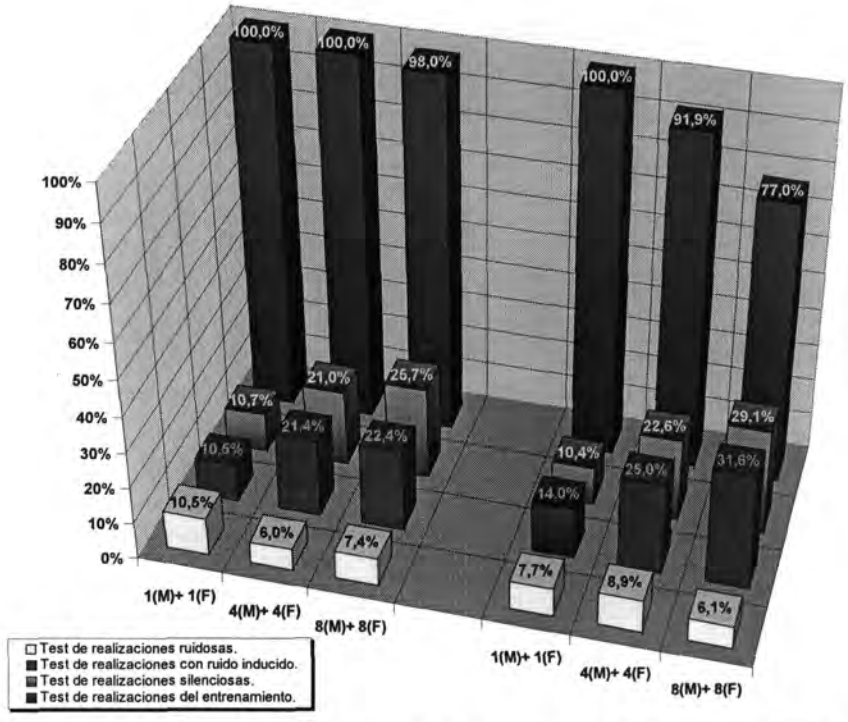

Figura A.134. Tasas de reconocimiento para el experimento PLP-12b (RASTA-P-Cepstrum [10 coeficientes]). La parte de la izquierda corresponde a los entrenamientos efectuados solamente con realizaciones silenciosas.

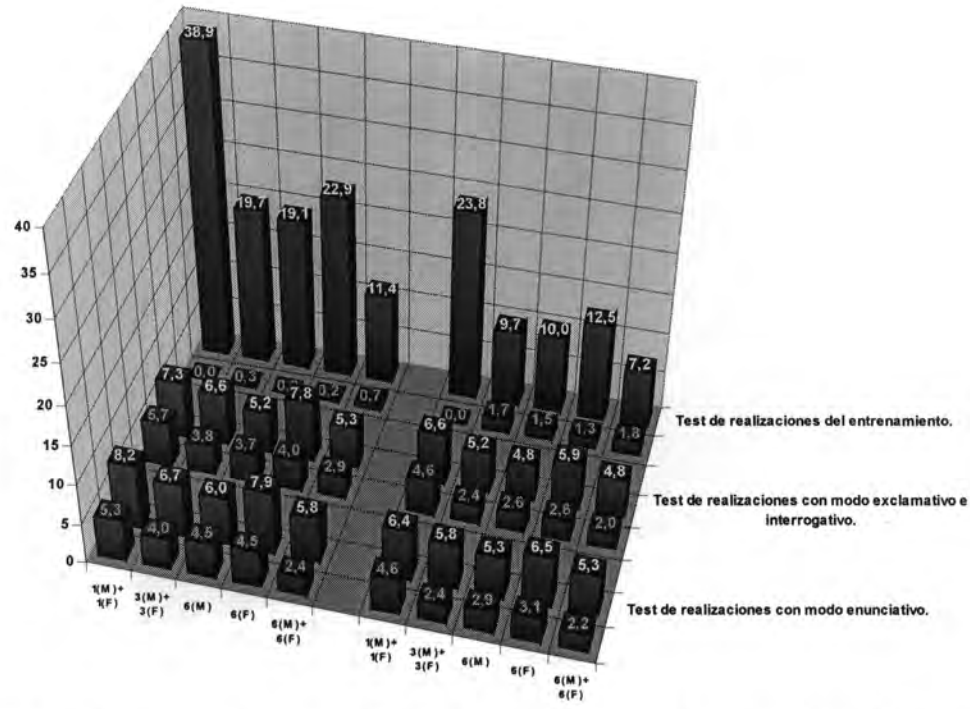

Figura A.135. Tasas de fiabilidad para el experimento PLP-12a (RASTA-P-Cepstrum [10 coeficientes]). En verde aparecen las tasas para los casos reconocidos correctamente. En rojo los casos clasificados erróneamente.

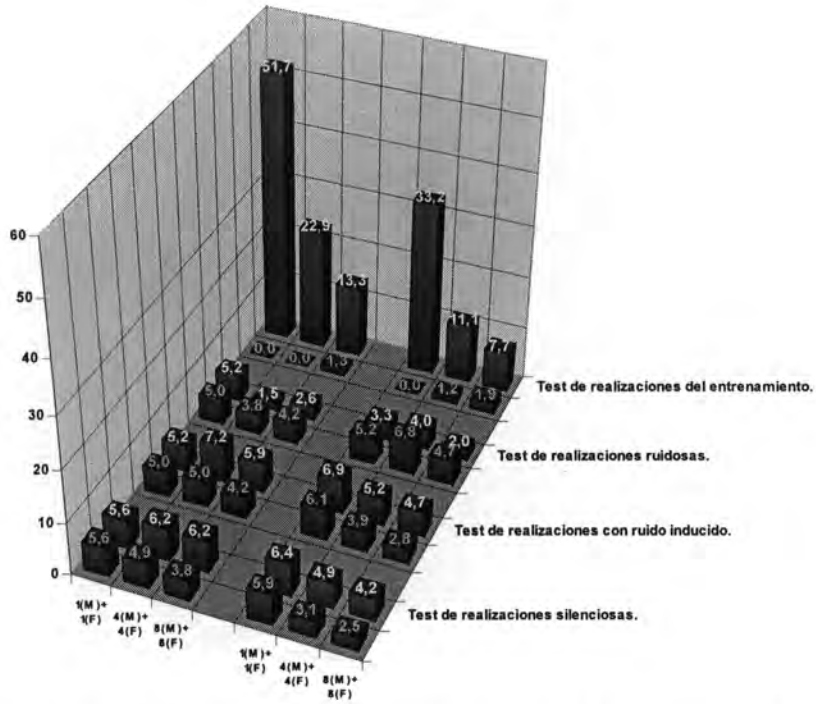

Figura A.136. Tasas de fiabilidad para el experimento PLP-12b (RASTA-P-Cepstrum [10 coeficientes]). En verde aparecen las tasas para los casos reconocidos correctamente. En rojo los casos clasificados erróneamente.

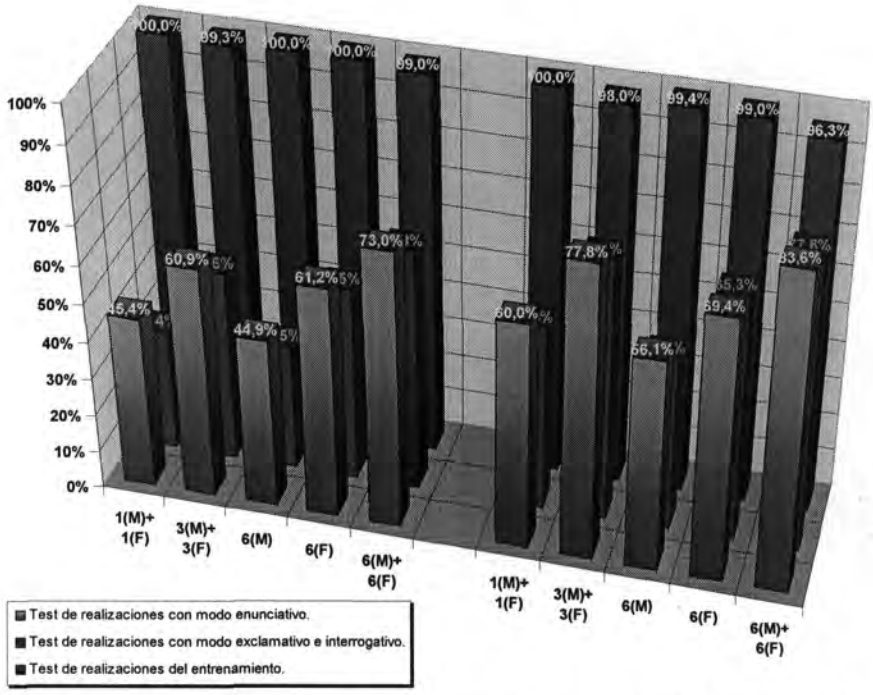

Figura A.137. Tasas de reconocimiento para el experimento PLP-13a (RASTA-P-Cepstrum [10 coeficientes + 10 coeficientes delta <5 tramas>]). La parte de la izquierda corresponde a los entrenamientos efectuados solamente con realizaciones con modo enunciativo.

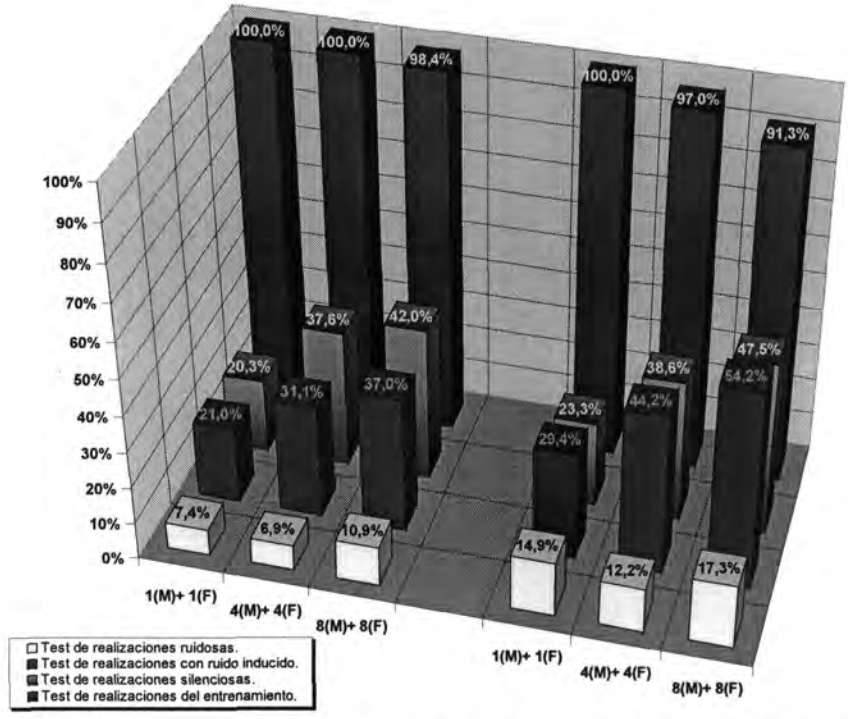

Figura A.138. Tasas de reconocimiento para el experimento PLP-13b (RASTA-P-Cepstrum [10 coeficientes + 10 coeficientes delta <5 tramas>]). La parte de la izquierda corresponde a los entrenamientos efectuados solamente con realizaciones silenciosas.

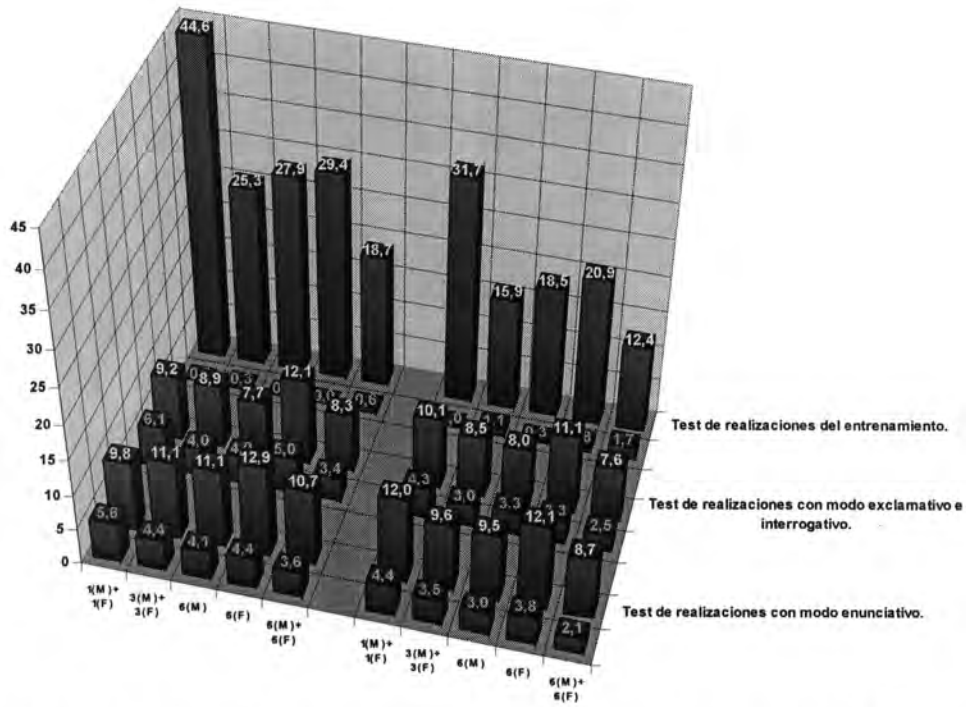

Figura A.139. Tasas de fiabilidad para el experimento PLP-13a (RASTA-P-Cepstrum [10 coeficientes + 10 coeficientes delta <5 tramas>]). En verde aparecen las tasas para los casos reconocidos correctamente. En rojo los casos clasificados erróneamente.

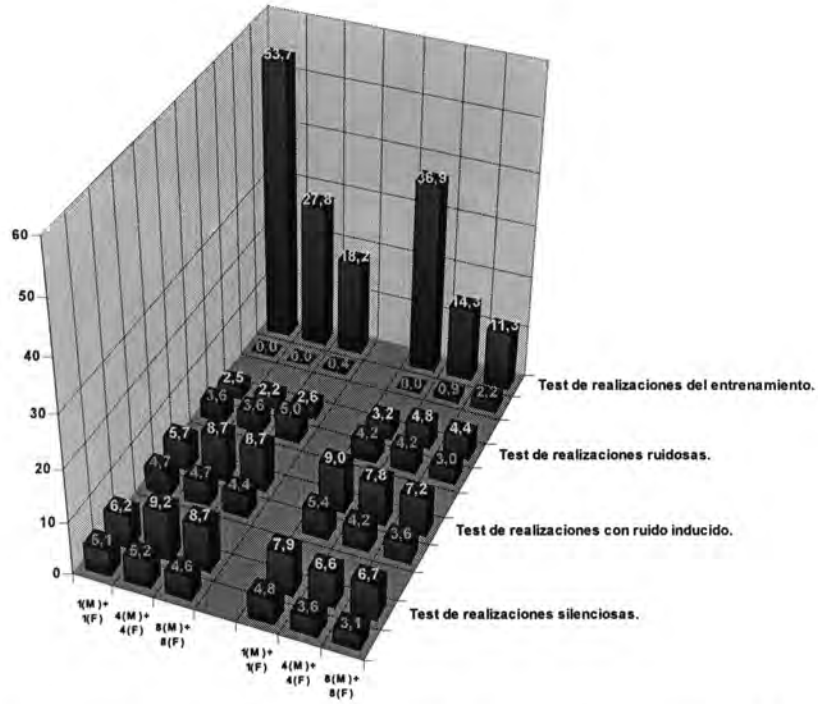

Figura A.140. Tasas de fiabilidad para el experimento PLP-13b (RASTA-P-Cepstrum [10 coeficientes + 10 coeficientes delta <5 tramas>]). En verde aparecen las tasas para los casos reconocidos correctamente. En rojo los casos clasificados erróneamente.

APÉNDICE B. RESULTADOS DE RECONOCIMIENTO POR PALABRA

B.1. PARAMETRIZADORES BASADOS EN LPC

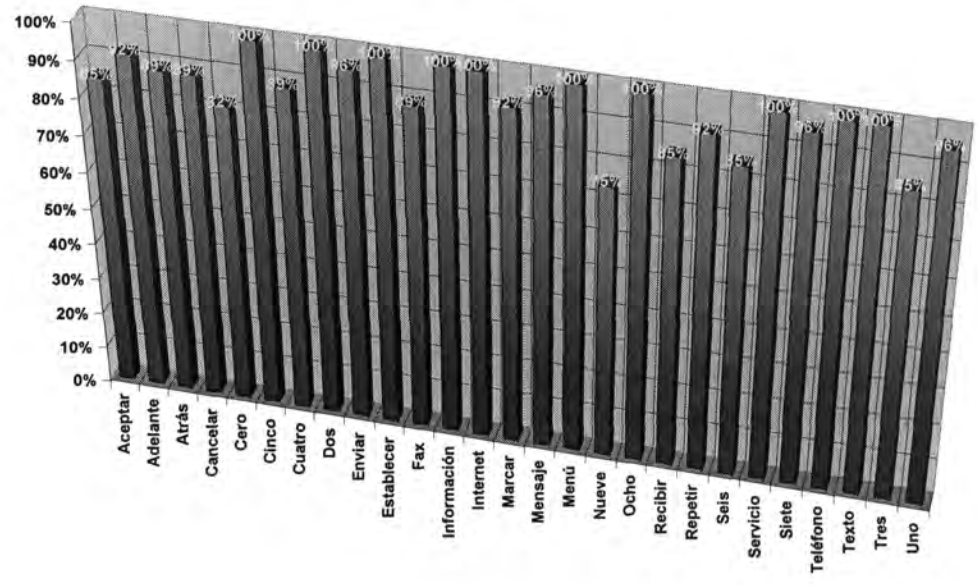

Figura B.1. Tasas de reconocimiento del test de realizaciones con entonación enunciativa del experimento LPC-6a (LPC-Bandas Energía [20 bandas + 20 bandas delta <5 tramas>]).

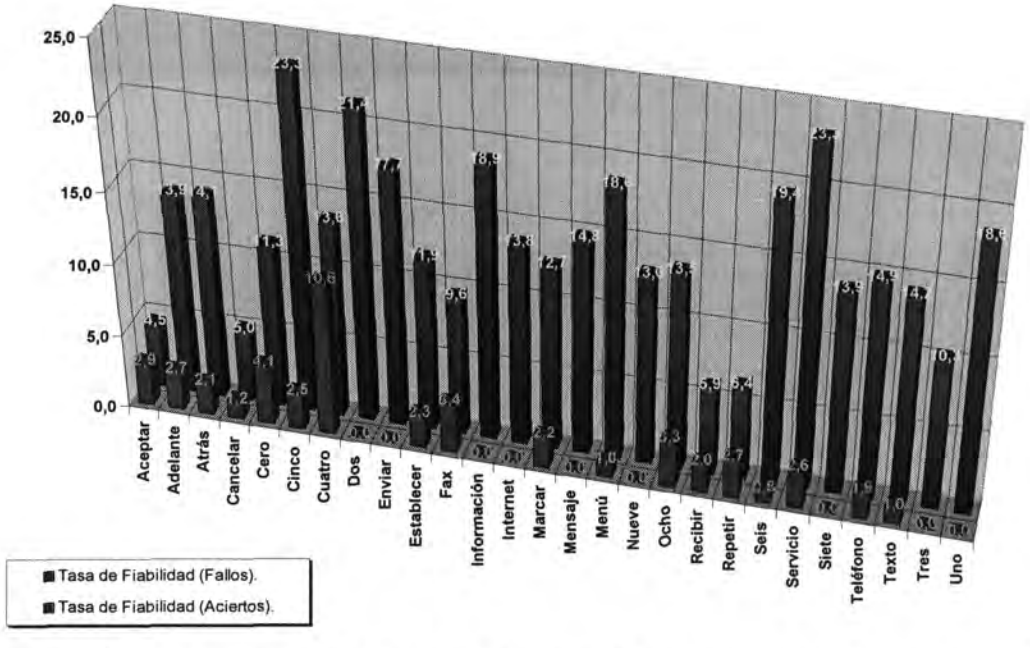

Figura B.2. Tasas de fiabilidad del test de realizaciones con entonación enunciativa del experimento LPC-6a (LPC-Bandas Energía [20 bandas + 20 bandas delta <5 tramas>]).

	Aceptar	Adelante	Atrás	Cancelar	Cero	Cinco	Cuatro	Dos	Enviar	Establecer	Fax	Información	Internet	Marcar	Mensaje	Menú	Nueve	Ocho	Recibir	Repetir	Seis	Servicio	Siete	Teléfono	Texto	Tres	Uno
Aceptar	85%	3%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	10%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Adelante	3%	92%	0%	3%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Atrás	0%	0%	89%	0%	0%	0%	0%	0%	0%	0%	3%	0%	0%	7%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Cancelar	7%	3%	0%	89%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Cero	0%	0%	0%	0%	82%	0%	3%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	10%	0%	0%	3%	0%	0%
Cinco	0%	0%	0%	0%	0%	100%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Cuatro	0%	0%	0%	0%	0%	0%	89%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	7%	0%	0%	0%	0%	0%	3%	0%	0%	0%
Dos	0%	0%	0%	0%	0%	0%	0%	100%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Enviar	0%	0%	0%	3%	0%	0%	0%	0%	96%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Establecer	0%	0%	0%	0%	0%	0%	0%	0%	0%	100%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Fax	0%	0%	10%	0%	0%	0%	0%	0%	0%	0%	89%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Información	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	100%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Internet	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	100%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Marcar	3%	0%	3%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	92%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Mensaje	0%	0%	0%	0%	3%	0%	0%	0%	0%	0%	0%	0%	0%	0%	96%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Menú	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	100%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Nueve	0%	3%	0%	0%	0%	0%	0%	0%	0%	7%	0%	0%	0%	0%	0%	3%	75%	0%	3%	7%	0%	0%	0%	0%	0%	0%	0%
Ocho	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	100%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Recibir	0%	0%	0%	0%	0%	0%	0%	0%	0%	3%	0%	0%	0%	0%	0%	0%	0%	0%	85%	10%	0%	0%	0%	0%	0%	0%	0%
Repetir	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	3%	92%	0%	3%	0%	0%	0%	0%	0%
Seis	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	7%	0%	85%	7%	0%	0%	0%	0%	0%
Servicio	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	100%	0%	0%	0%	0%	0%
Siete	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	3%	0%	0%	96%	0%	0%	0%	0%
Teléfono	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	100%	0%	0%	0%
Texto	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	100%	0%	0%
Tres	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	7%	0%	0%	3%	3%	85%	0%
Uno	0%	0%	0%	0%	0%	3%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	96%

Tabla B.1. Ejemplo de Tabla B.de confusión del test de realizaciones con entonación enunciativa del experimento LPC-6a (LPC-Bandas Energía [20 bandas + 20 bandas delta <5 tramas>]).

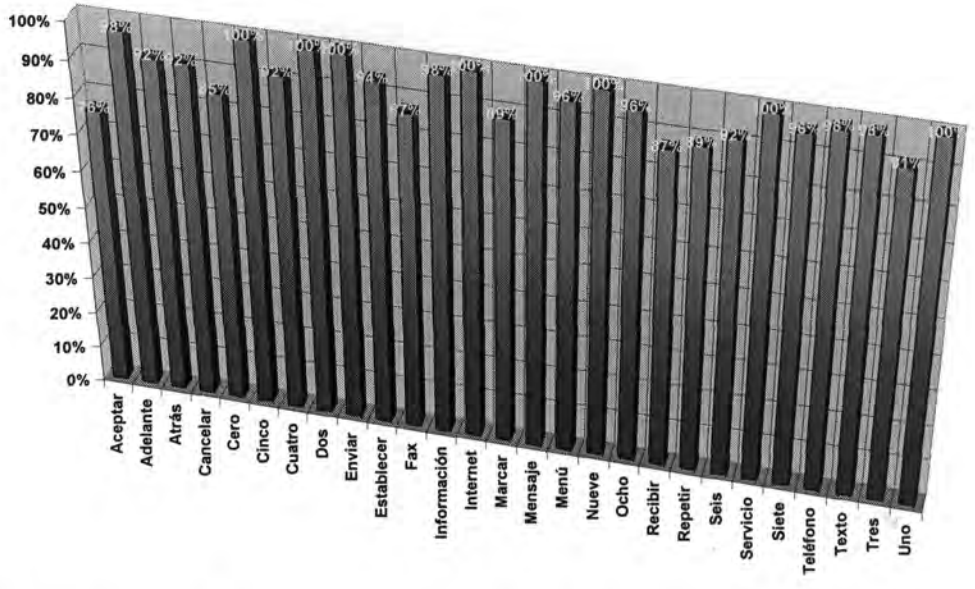

Figura B.3. Tasas de reconocimiento del test de realizaciones con entonación exclamativa/interrogativa del experimento LPC-6a (LPC-Bandas Energía [20 bandas + 20 bandas delta <5 tramas>]).

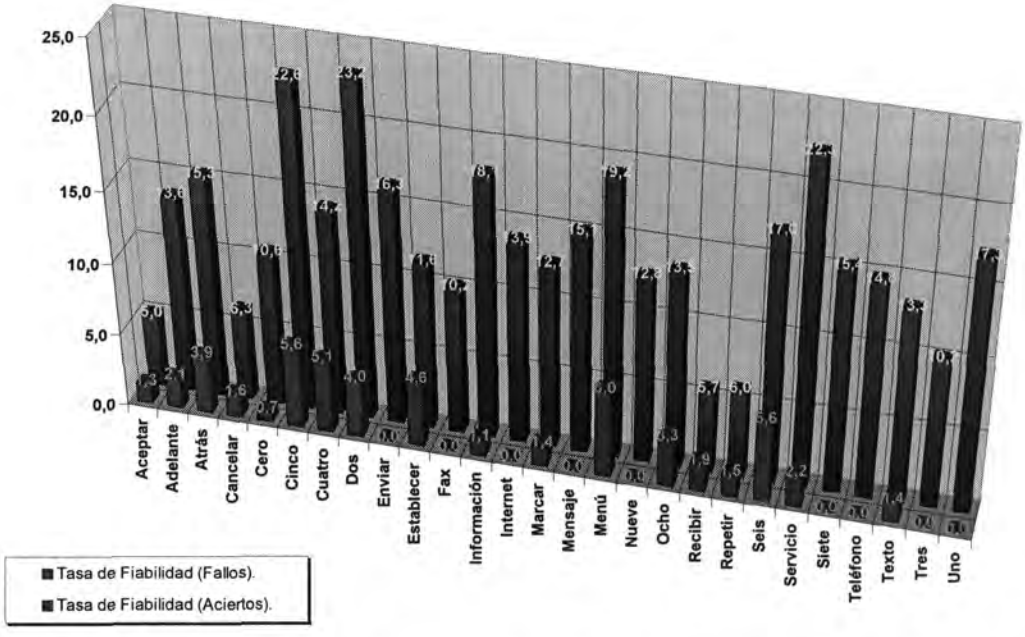

Figura B.4. Tasas de fiabilidad del test de realizaciones con entonación exclamativa/interrogativa del experimento LPC-6a (LPC-Bandas Energía [20 bandas + 20 bandas delta <5 tramas>]).

	Aceptar	Adelante	Atrás	Cancelar	Cero	Cinco	Cuatro	Dos	Enviar	Establecer	Fax	Información	Internet	Marcar	Mensaje	Menú	Nueve	Ocho	Recibir	Repetir	Seis	Servicio	Siete	Teléfono	Texto	Tres	Uno
Aceptar	76%	0%	1%	14%	1%	0%	0%	0%	0%	1%	0%	0%	0%	3%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Adelante	0%	98%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	1%	0%	0%	0%	0%	0%	0%	0%
Atrás	1%	0%	92%	3%	0%	0%	0%	1%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Cancelar	3%	1%	0%	92%	0%	0%	0%	0%	0%	0%	0%	1%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Cero	1%	0%	0%	0%	85%	5%	0%	0%	0%	0%	0%	0%	0%	0%	0%	1%	0%	0%	0%	0%	0%	5%	0%	0%	0%	0%	0%
Cinco	0%	0%	0%	0%	0%	100%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Cuatro	0%	0%	0%	0%	0%	0%	92%	0%	0%	0%	0%	3%	0%	0%	0%	0%	0%	3%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Dos	0%	0%	0%	0%	0%	0%	0%	100%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Enviar	0%	0%	0%	0%	0%	0%	0%	0%	100%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Establecer	0%	3%	0%	0%	0%	0%	0%	0%	0%	94%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	1%	0%	0%
Fax	0%	0%	10%	0%	0%	0%	0%	0%	0%	0%	87%	0%	0%	1%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Información	1%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	98%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Internet	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	100%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Marcar	3%	0%	0%	5%	0%	0%	1%	0%	0%	0%	0%	0%	0%	89%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Mensaje	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	100%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Menú	0%	0%	0%	0%	3%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	96%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Nueve	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	100%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Ocho	0%	0%	0%	0%	0%	1%	0%	1%	0%	0%	0%	0%	0%	0%	0%	0%	0%	96%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Recibir	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	87%	10%	0%	1%	0%	0%	0%	0%	0%
Repetir	0%	1%	0%	1%	0%	0%	0%	0%	0%	1%	0%	0%	0%	0%	0%	0%	0%	0%	5%	89%	0%	0%	0%	0%	0%	0%	0%
Seis	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	3%	1%	92%	1%	0%	0%	0%	0%	0%
Servicio	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	100%	0%	0%	0%	0%	0%
Siete	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	1%	1%	0%	0%	96%	0%	0%	0%	0%
Teléfono	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	98%	1%	0%	0%
Texto	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	1%	0%	0%	98%	0%	0%
Tres	0%	0%	1%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	1%	0%	5%	0%	0%	0%	0%	91%	0%
Uno	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	100%

Tabla B.2. Ejemplo de Tabla B.de confusión del test de realizaciones con entonaciones exclamativa e interrogativa del experimento LPC-6a (LPC-Bandas Energía [20 bandas + 20 bandas delta <5 tramas>]).

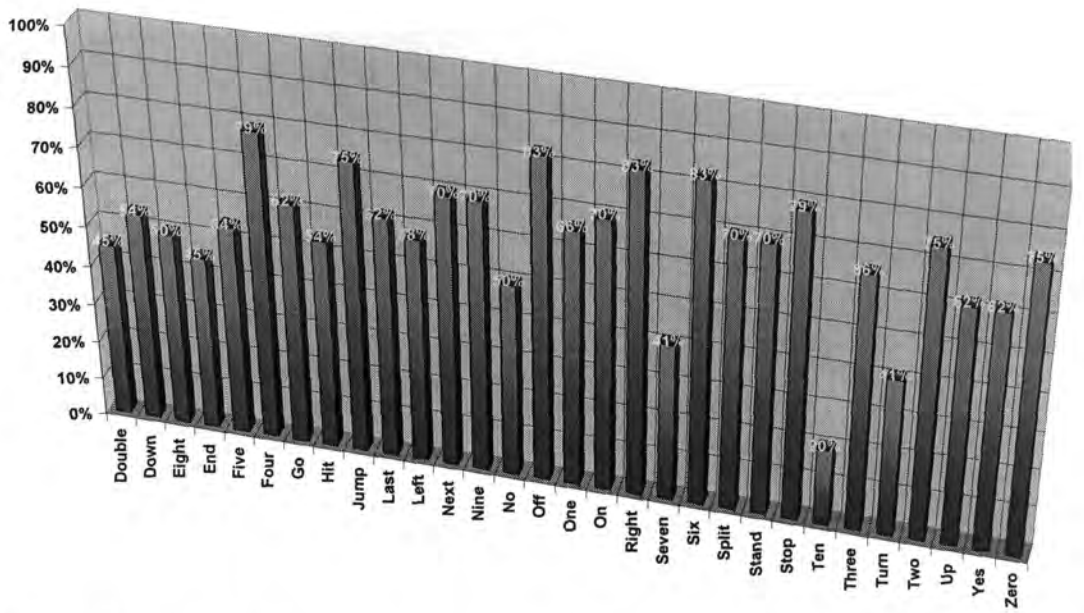

Figura B.5. Tasas de reconocimiento del test de realizaciones silenciosas del experimento LPC-6b (LPC-Bandas Energía [20 bandas + 20 bandas delta <5 tramas>]).

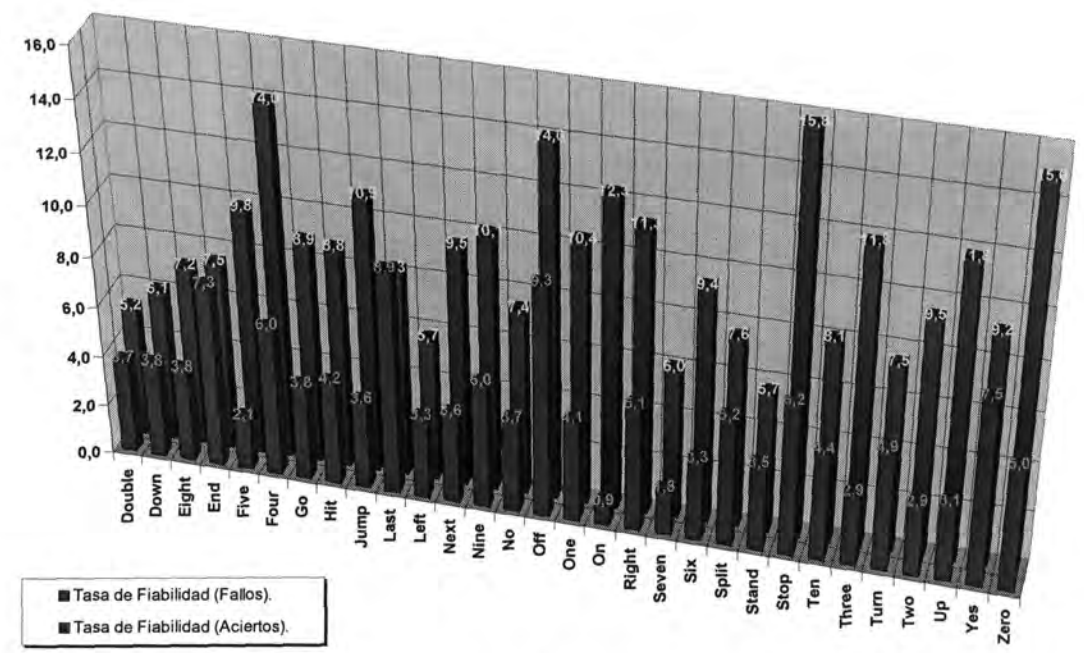

Figura B.6. Tasas de fiabilidad del test de realizaciones silenciosas del experimento LPC-6b (LPC-Bandas Energía [20 bandas + 20 bandas delta <5 tramas>]).

	Double	Down	Eight	End	Five	Four	Go	Hit	Jump	Last	Left	Next	Nine	No	Off	One	On	Right	Seven	Six	Split	Stand	Stop	Ten	Three	Turn	Two	Up	Yes	Zero
Double	45%	29%	0%	0%	0%	0%	12%	0%	4%	0%	0%	0%	0%	4%	0%	0%	0%	0%	0%	0%	0%	0%	0%	4%	0%	0%	0%	0%	0%	0%
Down	29%	54%	0%	0%	0%	0%	4%	0%	0%	0%	0%	0%	0%	4%	0%	0%	4%	4%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Eight	0%	0%	50%	0%	0%	0%	0%	12%	4%	0%	0%	4%	8%	0%	0%	0%	0%	0%	0%	4%	8%	0%	0%	4%	0%	4%	0%	0%	0%	0%
End	0%	0%	4%	45%	0%	0%	0%	0%	4%	0%	0%	0%	0%	0%	0%	4%	0%	8%	0%	0%	0%	4%	0%	16%	0%	0%	8%	0%	0%	4%
Five	4%	8%	0%	0%	54%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	4%	4%	16%	0%	0%	0%	4%	4%	0%	0%	0%	0%	0%	0%	0%
Four	4%	4%	0%	0%	0%	79%	8%	0%	0%	0%	0%	0%	0%	0%	0%	0%	4%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Go	29%	0%	0%	0%	0%	0%	62%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	4%	0%	0%	4%
Hit	0%	0%	8%	4%	4%	0%	0%	54%	0%	0%	0%	4%	0%	0%	0%	0%	0%	0%	0%	0%	8%	0%	0%	0%	12%	0%	0%	4%	0%	0%
Jump	0%	20%	0%	0%	0%	0%	0%	0%	75%	0%	0%	0%	0%	0%	0%	0%	0%	0%	4%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Last	0%	0%	0%	0%	0%	0%	0%	0%	8%	62%	8%	0%	0%	4%	0%	0%	0%	0%	0%	0%	0%	0%	4%	0%	0%	0%	4%	4%	4%	0%
Left	0%	0%	0%	4%	0%	0%	0%	4%	0%	4%	58%	12%	0%	0%	0%	0%	0%	0%	0%	0%	4%	0%	0%	4%	0%	4%	0%	4%	0%	0%
Next	4%	0%	0%	4%	0%	0%	0%	0%	0%	4%	8%	70%	0%	0%	0%	0%	0%	8%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Nine	0%	4%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	70%	4%	0%	4%	0%	0%	0%	0%	0%	0%	0%	0%	0%	16%	0%	0%	0%	0%
No	4%	0%	0%	0%	0%	0%	37%	0%	4%	0%	0%	0%	0%	50%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	4%
Off	0%	0%	0%	0%	0%	4%	0%	0%	0%	0%	0%	0%	0%	0%	83%	0%	0%	0%	0%	0%	0%	0%	4%	0%	0%	0%	0%	8%	0%	0%
One	8%	4%	0%	0%	4%	0%	4%	0%	0%	0%	0%	0%	0%	0%	0%	66%	8%	0%	0%	0%	0%	4%	0%	0%	0%	0%	0%	0%	0%	0%
On	0%	4%	0%	0%	0%	4%	0%	0%	0%	0%	4%	0%	0%	0%	0%	12%	70%	0%	0%	0%	0%	0%	0%	0%	0%	4%	0%	0%	0%	0%
Right	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	8%	0%	0%	4%	0%	83%	0%	0%	0%	0%	4%	0%	0%	0%	0%	0%	0%	0%
Seven	0%	4%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	4%	0%	0%	0%	0%	0%	41%	0%	0%	16%	0%	8%	0%	20%	0%	0%	0%	4%
Six	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	83%	4%	4%	4%	0%	4%	0%	0%	0%	0%	0%
Split	4%	0%	0%	0%	0%	0%	0%	12%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	4%	0%	70%	8%	0%	0%	0%	0%	0%	0%	0%	0%
Stand	0%	12%	0%	0%	4%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	8%	0%	0%	70%	4%	0%	0%	0%	0%	0%	0%	0%
Stop	4%	0%	0%	0%	0%	4%	4%	0%	0%	0%	0%	0%	0%	0%	4%	0%	0%	0%	4%	0%	0%	0%	79%	0%	0%	0%	0%	0%	0%	0%
Ten	4%	0%	4%	33%	0%	0%	0%	0%	0%	0%	0%	0%	4%	0%	0%	0%	0%	0%	4%	0%	0%	0%	0%	20%	0%	12%	12%	0%	0%	4%
Three	4%	0%	0%	0%	0%	0%	0%	8%	0%	0%	0%	0%	4%	0%	0%	0%	0%	4%	0%	4%	4%	0%	0%	0%	66%	4%	0%	0%	0%	0%
Turn	4%	12%	0%	4%	0%	0%	4%	0%	4%	0%	0%	0%	0%	0%	0%	0%	4%	0%	8%	0%	0%	4%	0%	12%	0%	41%	0%	0%	0%	0%
Two	8%	0%	0%	0%	0%	4%	4%	0%	4%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	4%	0%	0%	0%	75%	0%	0%	0%
Up	4%	0%	0%	0%	0%	0%	4%	0%	0%	0%	0%	0%	0%	0%	8%	0%	4%	0%	0%	0%	0%	0%	12%	0%	0%	4%	0%	62%	0%	0%
Yes	4%	0%	0%	0%	0%	0%	0%	0%	0%	8%	16%	4%	0%	0%	0%	0%	0%	0%	0%	4%	0%	0%	0%	0%	0%	0%	0%	0%	62%	0%
Zero	0%	0%	0%	0%	0%	0%	0%	0%	4%	0%	0%	0%	0%	0%	0%	0%	0%	0%	4%	0%	0%	8%	4%	0%	0%	0%	4%	0%	0%	75%

Tabla B.3. Ejemplo de Tabla B.de confusión del test de realizaciones silenciosas del experimento LPC-6b (LPC-Bandas Energía [20 bandas + 20 bandas delta <5 tramas>]).

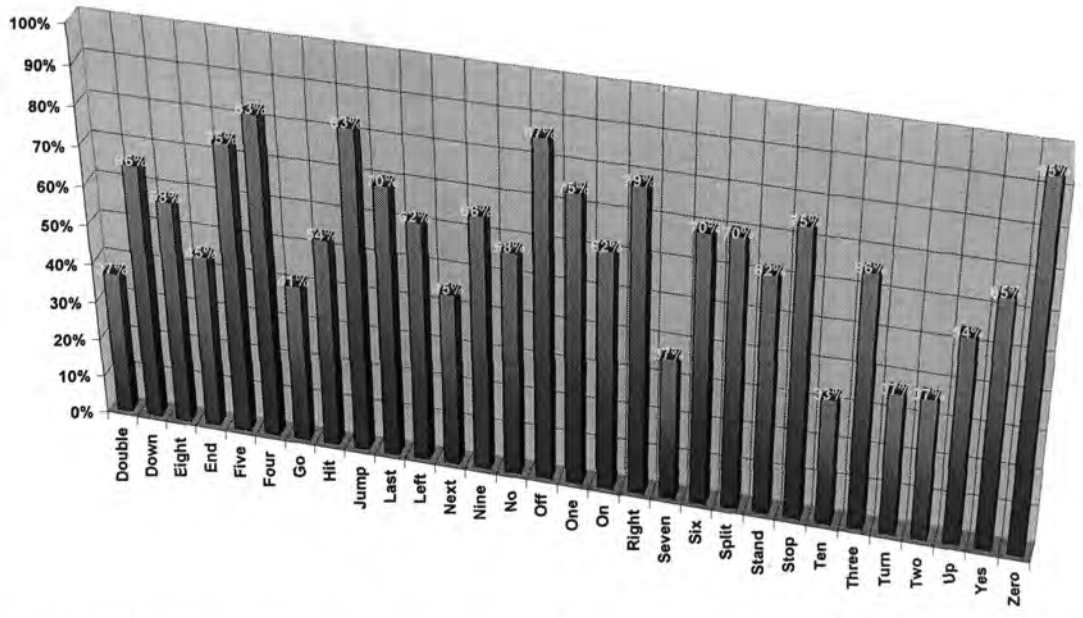

Figura B.7. Tasas de reconocimiento del test de realizaciones de habla con *stress* del experimento LPC-6b (LPC-Bandas Energía [20 bandas + 20 bandas delta <5 tramas>]).

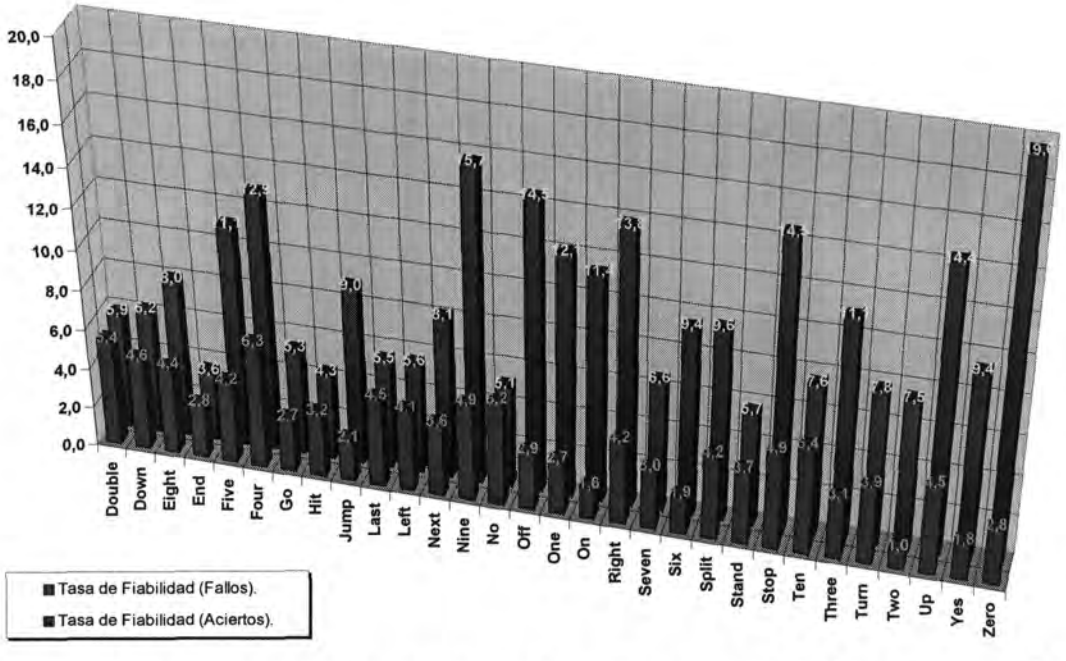

Figura B.8. Tasas de fiabilidad del test de realizaciones de habla con *stress* del experimento LPC-6b (LPC-Bandas Energía [20 bandas + 20 bandas delta <5 tramas>]).

	Double	Down	Eight	End	Five	Four	Go	Hit	Jump	Last	Left	Next	Nine	No	Off	One	On	Right	Seven	Six	Split	Stand	Stop	Ten	Three	Turn	Two	Up	Yes	Zero
Double	37%	33%	0%	0%	0%	0%	12%	0%	0%	0%	0%	0%	4%	0%	4%	0%	0%	0%	8%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Down	8%	66%	0%	0%	0%	0%	8%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	4%	0%	0%	0%	4%	0%	0%	0%	4%	0%	0%	0%	4%
Eight	0%	0%	58%	4%	0%	0%	0%	4%	0%	0%	0%	4%	0%	0%	0%	0%	0%	4%	0%	8%	8%	0%	0%	0%	8%	0%	0%	0%	0%	0%
End	0%	0%	4%	45%	0%	0%	0%	0%	4%	0%	0%	8%	8%	0%	0%	0%	0%	0%	0%	0%	0%	12%	0%	12%	0%	4%	0%	0%	0%	0%
Five	0%	0%	0%	0%	75%	4%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	4%	12%	0%	0%	4%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Four	0%	0%	0%	0%	4%	83%	0%	0%	0%	0%	4%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	4%	0%	0%	0%	0%	4%	0%	0%
Go	50%	0%	0%	0%	0%	0%	41%	0%	0%	0%	0%	0%	0%	8%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Hit	0%	0%	12%	4%	0%	0%	0%	54%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	25%	4%	0%	0%	0%	0%	0%	0%	0%	0%
Jump	4%	4%	0%	0%	0%	0%	0%	0%	83%	0%	0%	0%	4%	0%	0%	0%	0%	0%	0%	0%	0%	4%	0%	0%	0%	0%	0%	0%	0%	0%
Last	0%	0%	0%	0%	0%	0%	0%	0%	0%	70%	8%	4%	4%	0%	0%	0%	0%	8%	0%	0%	4%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Left	0%	0%	0%	4%	0%	0%	0%	0%	4%	8%	62%	4%	0%	0%	0%	0%	0%	0%	0%	0%	12%	0%	0%	0%	0%	4%	0%	0%	0%	0%
Next	0%	0%	0%	0%	0%	0%	0%	4%	0%	8%	12%	45%	4%	0%	0%	0%	0%	0%	0%	12%	8%	0%	0%	0%	0%	0%	0%	0%	4%	0%
Nine	0%	4%	0%	0%	8%	0%	0%	0%	0%	0%	4%	4%	66%	4%	0%	8%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
No	29%	0%	0%	0%	0%	0%	4%	0%	0%	0%	0%	0%	0%	58%	0%	0%	0%	0%	0%	0%	0%	0%	4%	0%	0%	0%	0%	0%	0%	4%
Off	0%	0%	0%	0%	0%	0%	4%	0%	0%	0%	0%	0%	0%	0%	87%	0%	0%	0%	0%	0%	0%	0%	8%	0%	0%	0%	0%	0%	0%	0%
One	4%	0%	0%	0%	12%	4%	0%	0%	0%	0%	0%	0%	4%	0%	0%	75%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
On	4%	0%	0%	0%	4%	8%	0%	0%	0%	0%	0%	0%	0%	0%	0%	12%	62%	4%	4%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Right	0%	0%	0%	0%	4%	0%	0%	0%	0%	0%	0%	0%	16%	0%	0%	0%	0%	79%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Seven	0%	12%	0%	0%	4%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	37%	0%	0%	33%	0%	0%	0%	12%	0%	0%	0%	0%
Six	0%	0%	4%	0%	0%	0%	0%	4%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	70%	16%	0%	0%	0%	4%	0%	0%	0%	0%	0%
Split	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	4%	4%	0%	70%	8%	0%	0%	12%	0%	0%	0%	0%	0%
Stand	0%	0%	0%	0%	8%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	20%	0%	4%	62%	0%	0%	0%	4%	0%	0%	0%	0%
Stop	0%	4%	0%	0%	0%	4%	4%	0%	0%	0%	0%	0%	0%	0%	4%	0%	0%	0%	0%	0%	0%	0%	75%	0%	0%	0%	0%	4%	0%	4%
Ten	0%	4%	0%	16%	4%	0%	0%	4%	4%	0%	8%	0%	0%	4%	0%	0%	0%	0%	0%	0%	0%	8%	0%	33%	0%	8%	4%	0%	0%	0%
Three	0%	0%	0%	0%	8%	0%	0%	0%	0%	0%	0%	0%	16%	0%	0%	0%	0%	0%	0%	0%	8%	0%	0%	0%	66%	0%	0%	0%	0%	0%
Turn	8%	16%	0%	4%	0%	0%	0%	0%	4%	0%	0%	0%	0%	0%	0%	4%	0%	0%	4%	0%	0%	16%	0%	4%	0%	37%	0%	0%	0%	0%
Two	16%	8%	0%	0%	0%	8%	8%	0%	0%	0%	0%	0%	0%	4%	0%	0%	0%	0%	4%	0%	0%	0%	4%	4%	0%	4%	37%	0%	0%	0%
Up	4%	0%	0%	0%	0%	4%	0%	0%	0%	8%	0%	0%	0%	0%	4%	0%	4%	0%	0%	0%	0%	4%	12%	0%	0%	4%	0%	54%	0%	0%
Yes	0%	0%	0%	0%	0%	0%	0%	4%	4%	0%	4%	13%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	65%	8%
Zero	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	4%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	95%

Tabla B.4. Ejemplo de Tabla B.de confusión del test de realizaciones con stress del experimento LPC-6b (LPC-Bandas Energía [20 bandas + 20 bandas delta <5 tramas>]).

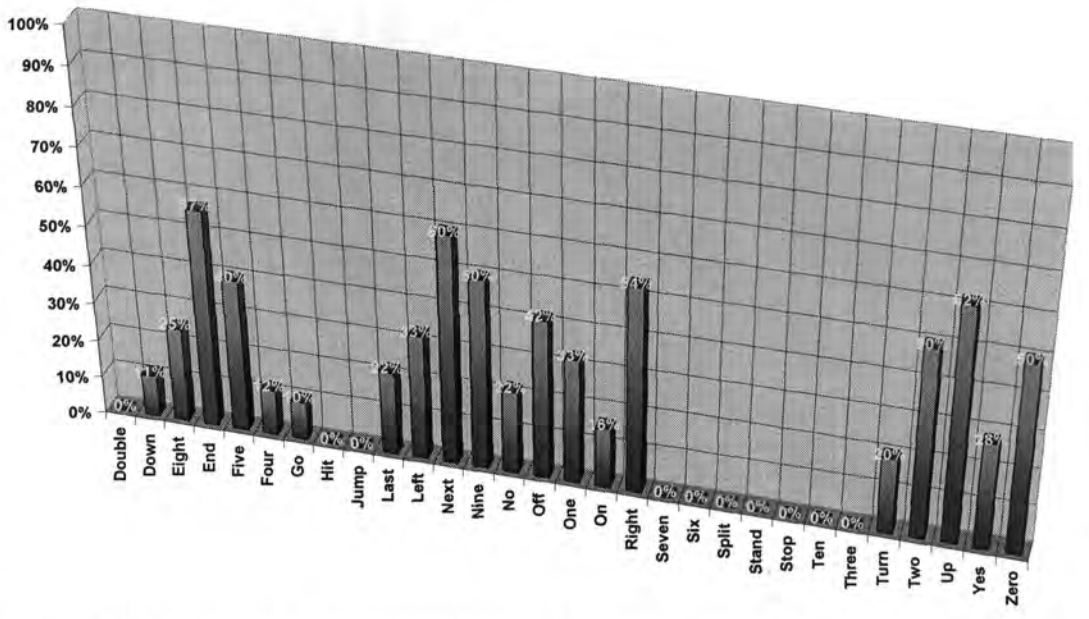

Figura B.9. Tasas de reconocimiento del test de realizaciones ruidosas del experimento LPC-12b (LPC-Bandas Energía Cepstrum [10 coeficientes + 10 coeficientes delta <5 tramas>]).

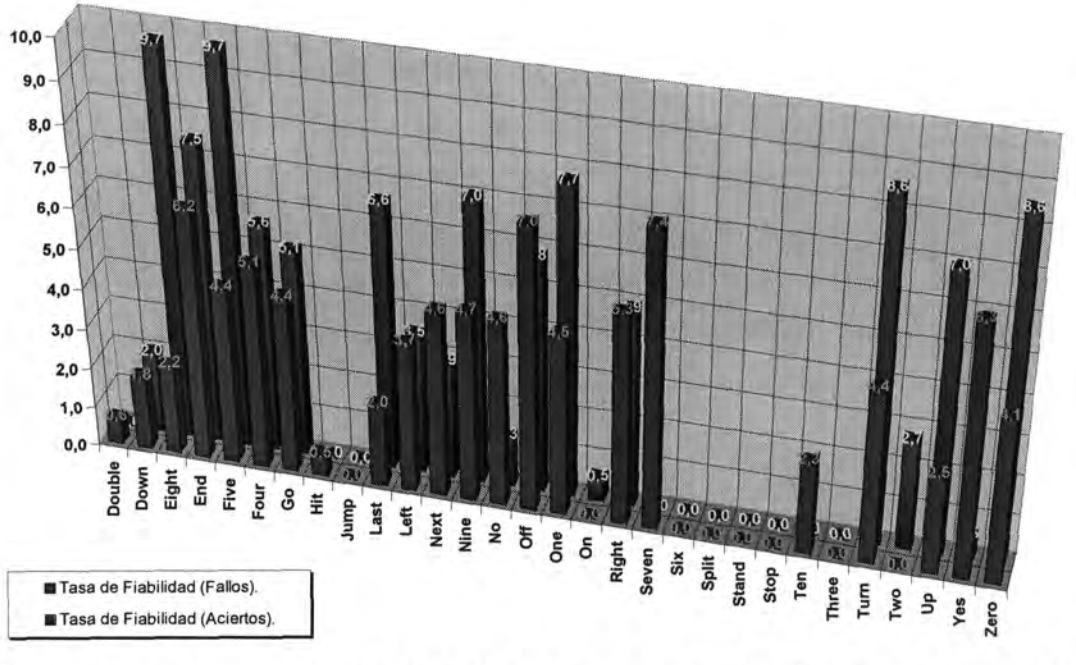

Figura B.10. Tasas de fiabilidad del test de realizaciones ruidosas del experimento LPC-12b (LPC-Bandas Energía Cepstrum [10 coeficientes + 10 coeficientes delta <5 tramas>]).

	Double	Down	Eight	End	Five	Four	Go	Hit	Jump	Last	Left	Next	Nine	No	Off	One	On	Right	Seven	Six	Split	Stand	Stop	Ten	Three	Turn	Two	Up	Yes	Zero
Double	0%	0%	0%	0%	0%	0%	44%	0%	0%	0%	0%	0%	11%	11%	0%	22%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	11%	0%	0%
Down	0%	11%	0%	0%	44%	0%	11%	0%	0%	0%	0%	0%	11%	11%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	11%	0%	0%
Eight	0%	0%	25%	50%	12%	0%	0%	12%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
End	0%	0%	0%	57%	14%	0%	0%	0%	0%	0%	0%	28%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Five	0%	0%	0%	0%	40%	0%	0%	0%	0%	0%	0%	0%	0%	0%	10%	0%	0%	50%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Four	0%	0%	0%	0%	12%	12%	0%	0%	0%	12%	0%	0%	0%	0%	25%	37%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Go	0%	0%	0%	0%	10%	20%	10%	0%	0%	0%	0%	0%	0%	0%	20%	30%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	10%	0%	0%
Hit	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	33%	33%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	33%	0%
Jump	0%	0%	0%	0%	30%	0%	20%	0%	0%	20%	0%	0%	0%	0%	0%	0%	0%	20%	0%	0%	0%	0%	0%	0%	0%	10%	0%	0%	0%	0%
Last	0%	0%	0%	0%	11%	0%	0%	0%	0%	22%	0%	11%	11%	0%	11%	0%	0%	22%	0%	0%	0%	0%	0%	0%	0%	0%	0%	11%	0%	0%
Left	0%	0%	0%	16%	0%	0%	0%	0%	0%	0%	33%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	33%	0%	0%	0%	0%	16%	0%
Next	0%	0%	20%	0%	0%	0%	0%	0%	0%	0%	20%	60%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Nine	0%	0%	0%	0%	12%	0%	0%	0%	0%	0%	12%	0%	50%	0%	0%	0%	0%	25%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
No	0%	0%	0%	0%	0%	22%	33%	0%	0%	0%	0%	0%	0%	22%	0%	11%	0%	11%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Off	0%	0%	0%	0%	28%	14%	0%	0%	0%	0%	0%	0%	0%	0%	42%	14%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
One	0%	0%	0%	0%	22%	0%	0%	0%	0%	11%	0%	0%	11%	0%	0%	33%	0%	11%	0%	0%	0%	0%	0%	0%	0%	0%	0%	11%	0%	0%
On	0%	0%	0%	0%	16%	0%	0%	0%	0%	0%	0%	0%	0%	0%	16%	33%	16%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	16%	0%	0%
Right	0%	0%	0%	0%	18%	0%	0%	0%	0%	9%	18%	0%	0%	0%	0%	0%	0%	54%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Seven	0%	0%	0%	28%	0%	0%	0%	0%	0%	0%	14%	14%	0%	0%	0%	0%	0%	14%	0%	0%	0%	0%	0%	0%	0%	28%	0%	0%	0%	0%
Six	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	100%
Split	0%	0%	0%	50%	0%	0%	0%	0%	0%	0%	0%	50%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Stand	0%	0%	0%	16%	0%	0%	0%	0%	0%	0%	0%	0%	16%	0%	0%	0%	0%	33%	16%	0%	0%	0%	0%	16%	0%	0%	0%	0%	0%	0%
Stop	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	33%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	66%	0%	0%
Ten	20%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	80%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Three	0%	0%	33%	0%	0%	0%	0%	0%	0%	0%	33%	33%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Turn	0%	0%	0%	10%	10%	0%	0%	0%	0%	10%	0%	0%	0%	0%	0%	0%	0%	30%	0%	0%	0%	0%	0%	0%	0%	20%	0%	10%	10%	0%
Two	25%	0%	0%	0%	0%	0%	25%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	50%	0%	0%	0%
Up	0%	0%	0%	0%	12%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	12%	0%	12%	0%	0%	0%	0%	0%	0%	0%	0%	0%	62%	0%	0%
Yes	0%	0%	0%	14%	0%	0%	0%	14%	0%	0%	0%	28%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	14%	0%	0%	0%	0%	28%	0%
Zero	0%	25%	0%	0%	0%	0%	25%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	50%

Tabla B.5. Ejemplo de Tabla B.de confusión del test de realizaciones ruidosas del experimento LPC-12b (LPC-Bandas Energía Cepstrum [10 coeficientes + 10 coeficientes delta <5 tramas>]).

B.2. PARAMETRIZADORES BASADOS EN FFT

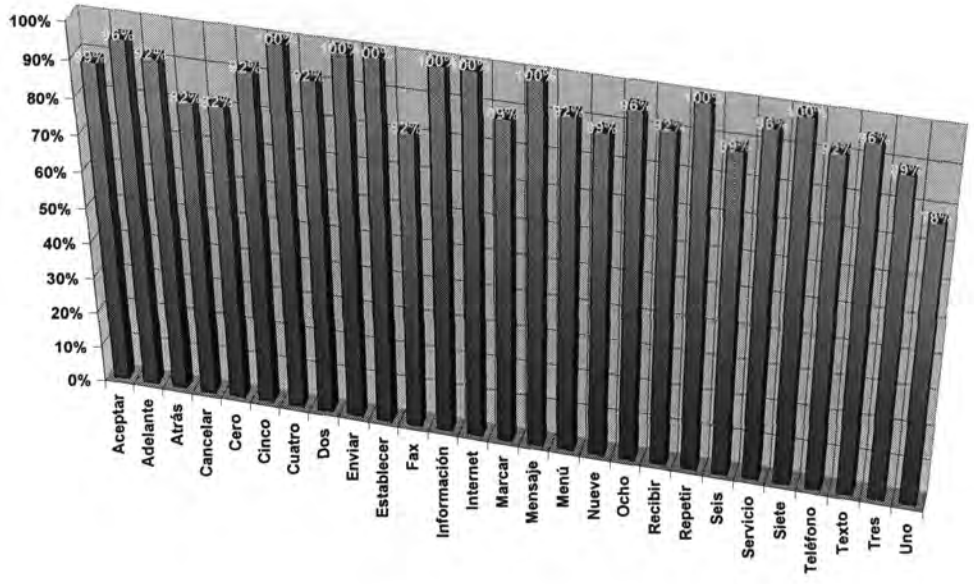

Figura B.11. Tasas de reconocimiento del test de realizaciones con entonación enunciativa del experimento FFT-2a (FFT-Bandas Energía [20 bandas + 20 bandas delta <5 tramas>]).

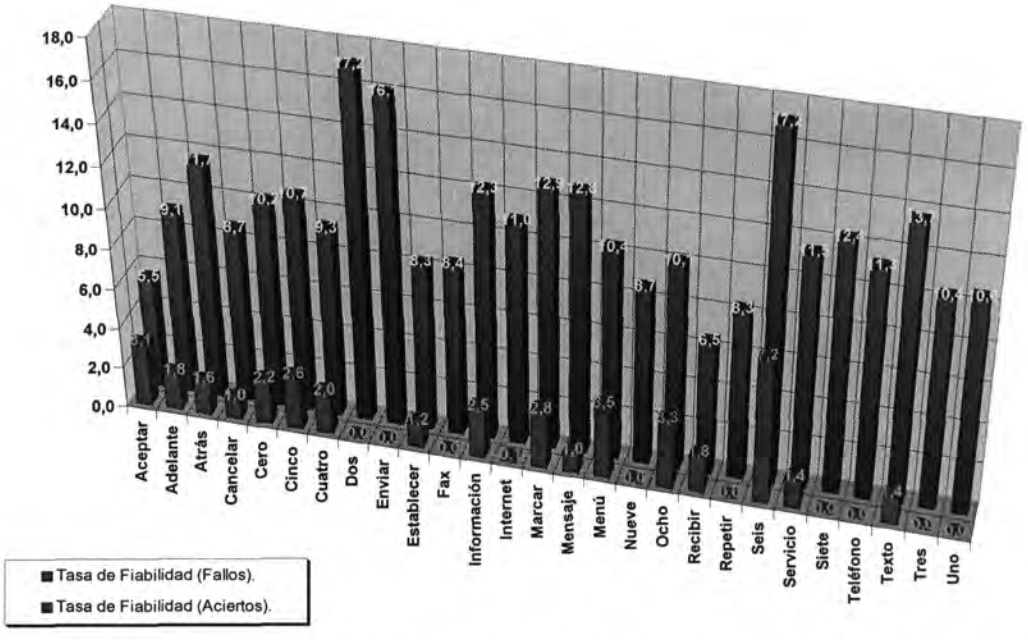

Figura B.12. Tasas de fiabilidad del test de realizaciones con entonación enunciativa del experimento FFT-2a (FFT-Bandas Energía [20 bandas + 20 bandas delta <5 tramas>]).

	Aceptar	Adelante	Atrás	Cancelar	Cero	Cinco	Cuatro	Dos	Enviar	Establecer	Fax	Información	Internet	Marcar	Mensaje	Menú	Nueve	Ocho	Recibir	Repetir	Seis	Servicio	Siete	Teléfono	Texto	Tres	Uno
Aceptar	89%	0%	0%	0%	0%	0%	0%	0%	0%	3%	0%	0%	0%	7%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Adelante	0%	96%	0%	0%	0%	0%	0%	0%	0%	0%	0%	3%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Atrás	0%	0%	92%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	7%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Cancelar	0%	3%	0%	82%	0%	0%	0%	0%	0%	7%	0%	3%	0%	3%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Cero	0%	0%	0%	0%	82%	7%	0%	0%	0%	3%	0%	0%	0%	0%	0%	0%	0%	0%	3%	0%	0%	3%	0%	0%	0%	0%	0%
Cinco	0%	0%	0%	0%	3%	92%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	3%	0%	0%
Cuatro	0%	0%	0%	0%	0%	0%	100%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Dos	0%	0%	3%	0%	0%	0%	0%	92%	0%	0%	0%	3%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Enviar	0%	0%	0%	0%	0%	0%	0%	0%	100%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Establecer	0%	0%	0%	0%	0%	0%	0%	0%	0%	100%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Fax	0%	0%	17%	0%	0%	0%	0%	0%	0%	0%	82%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Información	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	100%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Internet	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	100%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Marcar	7%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	3%	89%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Mensaje	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	100%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Menú	0%	3%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	92%	0%	0%	0%	0%	0%	3%	0%	0%	0%	0%	0%
Nueve	0%	3%	0%	0%	0%	0%	0%	0%	0%	3%	0%	0%	0%	0%	0%	3%	89%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Ocho	0%	0%	0%	0%	0%	0%	0%	0%	0%	3%	0%	0%	0%	0%	0%	0%	0%	96%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Recibir	0%	0%	0%	0%	0%	0%	0%	0%	0%	3%	0%	0%	0%	0%	3%	0%	0%	0%	92%	0%	0%	0%	0%	0%	0%	0%	0%
Repetir	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	100%	0%	0%	0%	0%	0%	0%	0%
Seis	0%	0%	0%	3%	3%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	89%	3%	0%	0%	0%	0%	0%
Servicio	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	3%	0%	0%	96%	0%	0%	0%	0%	0%
Siete	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	100%	0%	0%	0%	0%
Teléfono	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	7%	0%	0%	0%	0%	92%	0%	0%	0%
Texto	3%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	96%	0%	0%
Tres	0%	0%	3%	0%	0%	0%	0%	0%	0%	3%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	3%	0%	0%	0%	0%	89%	0%
Uno	0%	0%	0%	0%	0%	0%	7%	0%	0%	0%	0%	0%	0%	0%	0%	10%	0%	3%	0%	0%	0%	0%	0%	0%	0%	0%	78%

Tabla B.6. Ejemplo de Tabla B.de confusión del test de realizaciones con entonación enunciativa del experimento FFT-2a (FFT-Bandas Energía [20 bandas + 20 bandas delta <5 tramas>]).

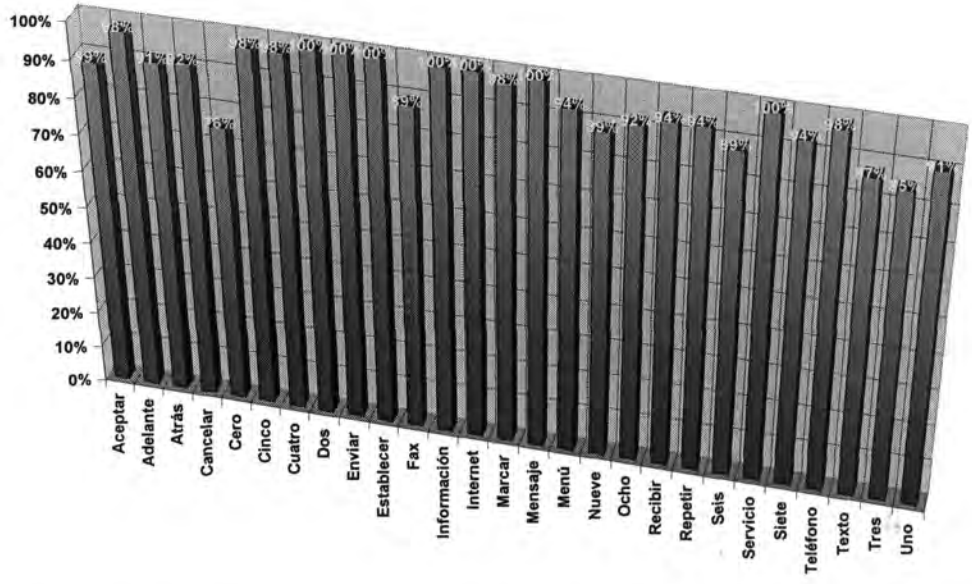

Figura B.13. Tasas de reconocimiento del test de realizaciones con entonación exclamativa/interrogativa del experimento FFT-5a (FFT-Bandas Energía Normalizadas [20 bandas + 20 bandas delta <5 tramas>]).

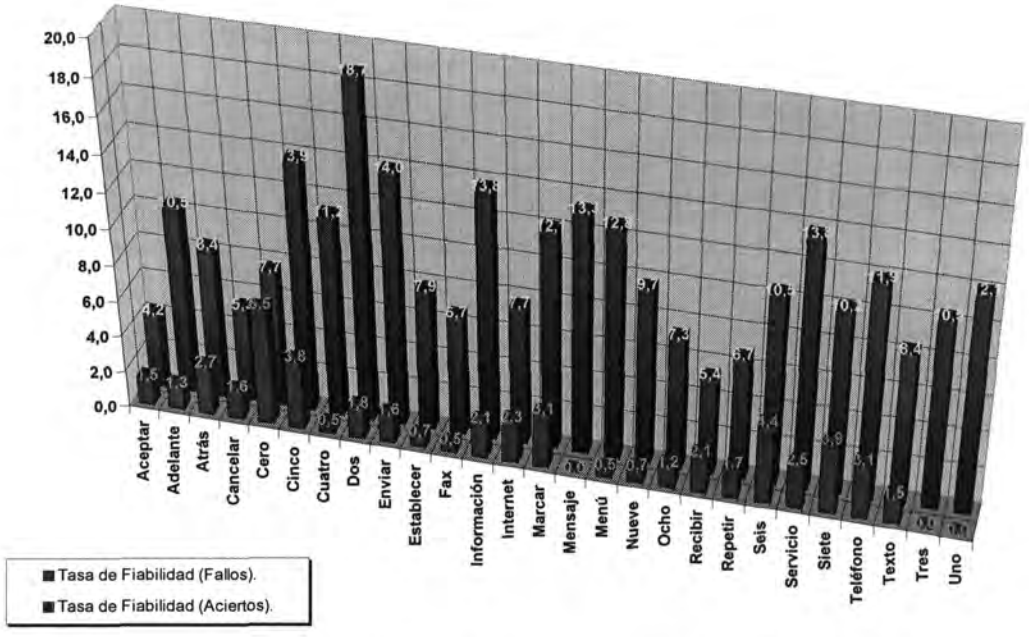

Figura B.14. Tasas de fiabilidad del test de realizaciones con entonación exclamativa/interrogativa del experimento FFT-5a (FFT-Bandas Energía Normalizadas [20 bandas + 20 bandas delta <5 tramas>]).

	Aceptar	Adelante	Atrás	Cancelar	Cero	Cinco	Cuatro	Dos	Enviar	Establecer	Fax	Información	Internet	Marcar	Mensaje	Menú	Nueve	Ocho	Recibir	Repetir	Seis	Servicio	Siete	Teléfono	Texto	Tres	Uno
Aceptar	89%	0%	0%	3%	0%	0%	0%	0%	0%	1%	1%	1%	0%	1%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Adelante	0%	98%	0%	0%	0%	0%	0%	0%	0%	0%	0%	1%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Atrás	3%	0%	91%	1%	0%	0%	0%	0%	0%	1%	0%	0%	0%	1%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Cancelar	5%	0%	0%	92%	0%	0%	0%	0%	0%	0%	0%	1%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Cero	0%	0%	0%	0%	76%	1%	0%	0%	0%	3%	0%	0%	0%	0%	0%	0%	0%	1%	0%	0%	0%	14%	0%	0%	1%	0%	0%
Cinco	0%	0%	0%	0%	0%	98%	0%	0%	0%	0%	0%	1%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Cuatro	0%	0%	0%	0%	0%	0%	98%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	1%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Dos	0%	0%	0%	0%	0%	0%	0%	100%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Enviar	0%	0%	0%	0%	0%	0%	0%	0%	100%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Establecer	0%	0%	0%	0%	0%	0%	0%	0%	0%	100%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Fax	0%	0%	8%	0%	0%	0%	0%	1%	0%	0%	89%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Información	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	100%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Internet	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	100%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Marcar	1%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	98%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Mensaje	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	100%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Menú	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	1%	1%	0%	0%	94%	1%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Nueve	0%	1%	0%	0%	0%	0%	0%	0%	0%	0%	0%	1%	7%	0%	0%	0%	89%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Ocho	1%	0%	0%	0%	0%	1%	3%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	92%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Recibir	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	94%	3%	0%	1%	0%	0%	0%	0%	0%
Repetir	0%	1%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	3%	94%	0%	0%	0%	0%	0%	0%	0%
Seis	0%	0%	0%	0%	0%	0%	0%	0%	1%	0%	0%	0%	0%	0%	0%	0%	0%	0%	5%	0%	89%	1%	1%	0%	0%	0%	0%
Servicio	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	100%	0%	0%	0%	0%	0%
Siete	0%	3%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	1%	0%	0%	0%	94%	0%	0%	0%	0%
Teléfono	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	1%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	98%	0%	0%	0%
Texto	1%	0%	0%	0%	1%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	5%	0%	3%	87%	0%	0%
Tres	0%	0%	1%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	1%	8%	0%	0%	0%	1%	85%	0%
Uno	0%	0%	0%	0%	0%	1%	0%	0%	0%	0%	0%	3%	1%	0%	0%	1%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	91%

Tabla B.7. Ejemplo de Tabla B.de confusión del test de realizaciones con entonaciones exclamativa e interrogativa del experimento FFT-5a (FFT-Bandas Energía Normalizadas [20 bandas + 20 bandas delta <5 tramas>]).

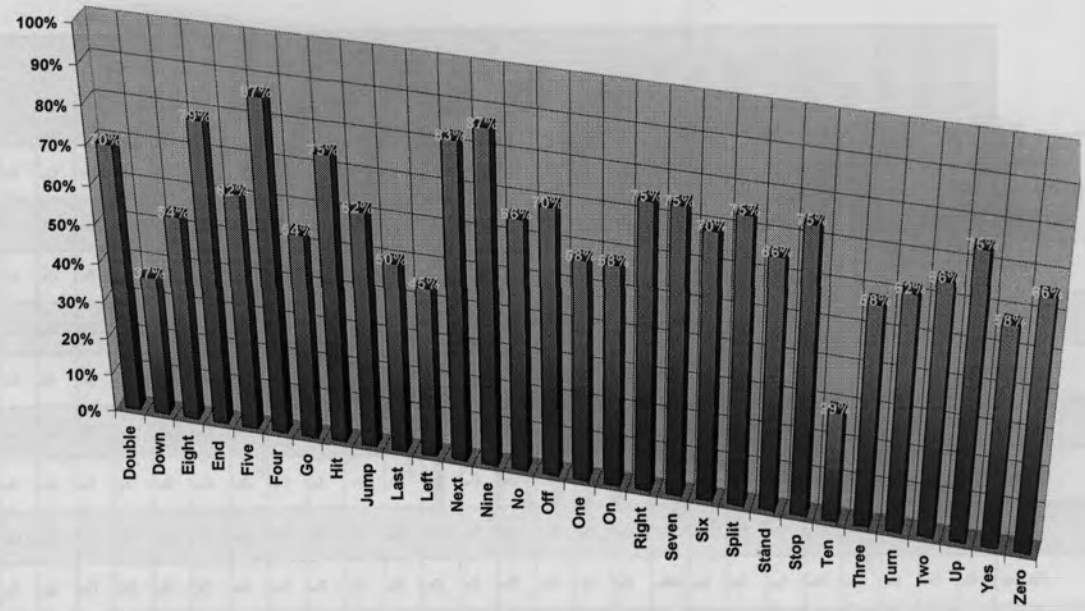

Figura B.15. Tasas de reconocimiento del test de realizaciones silenciosas del experimento FFT-2b (FFT-Bandas Energía [20 bandas + 20 bandas delta <5 tramas>]).

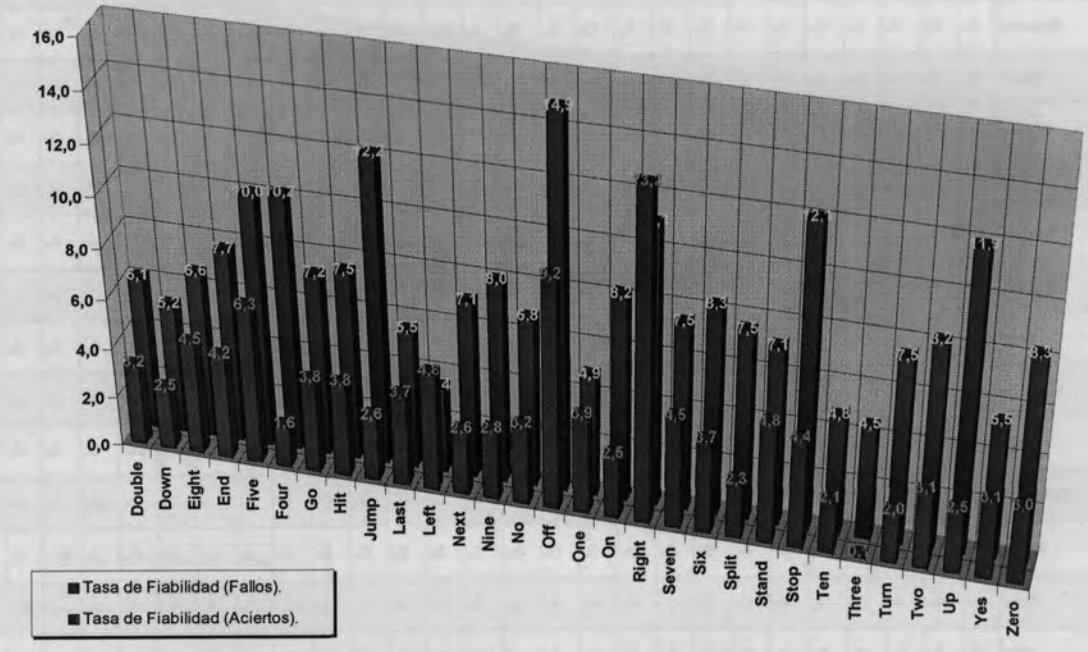

Figura B.16. Tasas de fiabilidad del test de realizaciones silenciosas del experimento FFT-2b (FFT-Bandas Energía [20 bandas + 20 bandas delta <5 tramas>]).

	Double	Down	Eight	End	Five	Four	Go	Hit	Jump	Last	Left	Next	Nine	No	Off	One	On	Right	Seven	Six	Split	Stand	Stop	Ten	Three	Turn	Two	Up	Yes	Zero
Double	70%	4%	0%	0%	0%	0%	8%	0%	4%	0%	0%	0%	8%	4%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Down	20%	37%	0%	0%	0%	0%	16%	0%	0%	0%	0%	0%	4%	8%	0%	0%	4%	0%	0%	0%	0%	0%	0%	0%	8%	0%	0%	0%	0%	0%
Eight	0%	0%	54%	25%	0%	0%	0%	0%	0%	0%	0%	4%	0%	0%	0%	4%	0%	0%	0%	0%	8%	0%	0%	4%	0%	0%	0%	0%	0%	0%
End	0%	0%	0%	79%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	4%	4%	0%	0%	0%	0%	8%	0%	4%	0%	0%	0%	0%	0%	0%
Five	0%	4%	0%	4%	62%	8%	0%	0%	0%	0%	0%	0%	0%	0%	0%	8%	0%	4%	0%	0%	0%	8%	0%	0%	0%	0%	0%	0%	0%	0%
Four	0%	0%	0%	0%	8%	87%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	4%
Go	20%	0%	0%	0%	0%	4%	54%	0%	0%	0%	0%	0%	0%	16%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	4%	0%	0%	0%	0%
Hit	0%	0%	4%	4%	4%	0%	0%	75%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	4%	4%	0%	0%	4%	0%	0%	0%	0%	0%
Jump	8%	0%	0%	0%	4%	0%	12%	0%	62%	0%	0%	0%	4%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	8%
Last	0%	0%	0%	0%	0%	0%	4%	0%	0%	50%	25%	4%	0%	4%	4%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	4%	4%	0%
Left	4%	0%	4%	4%	0%	0%	0%	0%	0%	4%	45%	12%	0%	8%	0%	0%	0%	0%	0%	0%	0%	4%	0%	0%	0%	0%	0%	0%	12%	0%
Next	0%	0%	0%	4%	0%	0%	0%	0%	0%	0%	8%	83%	0%	4%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Nine	4%	0%	0%	0%	0%	0%	4%	0%	0%	0%	0%	0%	87%	0%	0%	4%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
No	8%	4%	0%	0%	0%	0%	8%	0%	0%	0%	0%	0%	8%	66%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	4%	0%	0%
Off	4%	0%	0%	8%	4%	0%	0%	0%	0%	0%	0%	0%	0%	0%	70%	0%	4%	0%	0%	0%	0%	4%	0%	0%	0%	0%	0%	4%	0%	0%
One	0%	0%	0%	0%	4%	0%	4%	0%	0%	0%	0%	0%	29%	4%	0%	58%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
On	0%	4%	0%	4%	0%	0%	8%	0%	0%	0%	0%	0%	0%	0%	0%	20%	58%	0%	0%	0%	0%	0%	0%	0%	0%	4%	0%	0%	0%	0%
Right	0%	0%	0%	0%	0%	0%	4%	0%	0%	4%	4%	0%	8%	0%	0%	0%	0%	75%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	4%	0%
Seven	4%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	4%	0%	0%	75%	0%	0%	4%	0%	4%	0%	0%	0%	0%	0%	8%
Six	0%	0%	0%	0%	0%	0%	0%	12%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	70%	12%	0%	4%	0%	0%	0%	0%	0%	0%	0%
Split	0%	0%	0%	0%	0%	0%	0%	0%	4%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	12%	75%	8%	0%	0%	0%	0%	0%	0%	0%	0%
Stand	4%	4%	0%	4%	12%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	66%	4%	0%	0%	4%	0%	0%	0%	0%
Stop	0%	0%	0%	0%	0%	0%	0%	0%	4%	0%	0%	0%	0%	0%	8%	0%	0%	0%	0%	0%	4%	4%	75%	0%	0%	0%	0%	0%	0%	4%
Ten	4%	8%	0%	33%	0%	0%	8%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	29%	0%	4%	12%	0%	0%	0%
Three	0%	0%	0%	0%	0%	0%	0%	12%	0%	0%	0%	0%	8%	0%	0%	0%	0%	0%	0%	0%	8%	4%	0%	8%	58%	0%	0%	0%	0%	0%
Turn	0%	8%	0%	4%	0%	0%	4%	0%	0%	0%	0%	0%	4%	4%	0%	0%	4%	0%	0%	0%	0%	0%	0%	4%	0%	62%	4%	0%	0%	0%
Two	0%	0%	0%	0%	0%	8%	4%	4%	4%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	12%	0%	0%	66%	0%	0%	0%
Up	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	16%	0%	0%	0%	0%	0%	0%	0%	8%	0%	0%	0%	0%	75%	0%	0%
Yes	0%	0%	0%	0%	0%	0%	0%	0%	4%	4%	8%	25%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	58%	0%
Zero	4%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	4%	0%	0%	0%	0%	0%	8%	0%	0%	4%	4%	4%	0%	0%	4%	0%	0%	66%

Tabla B.8. Ejemplo de Tabla B.de confusión del test de realizaciones silenciosas del experimento FFT-2b (FFT-Bandas Energía [20 bandas + 20 bandas delta <5 tramas>]).

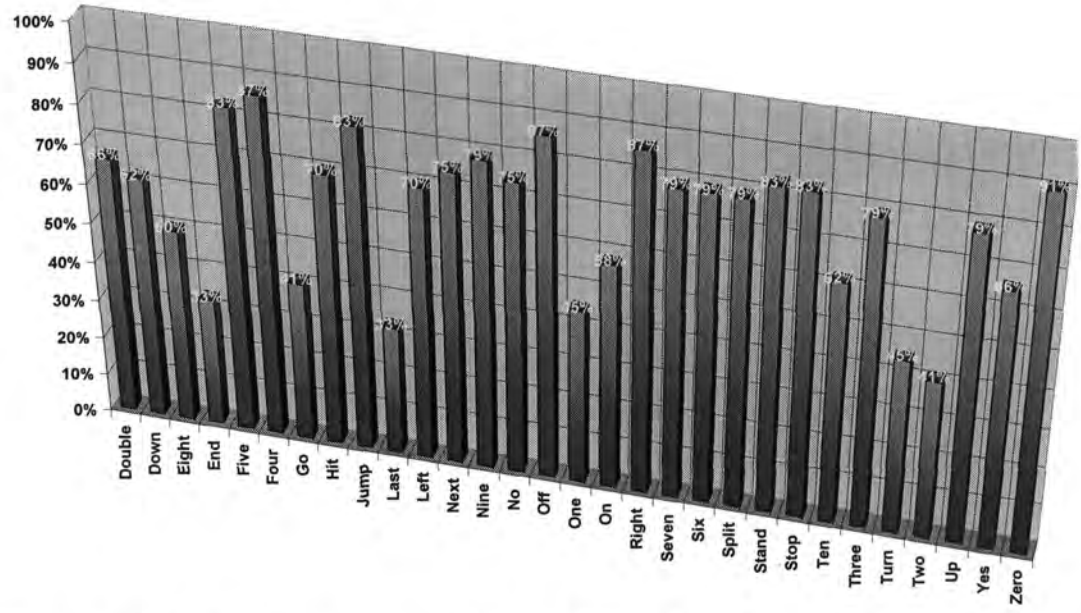

Figura B.17. Tasas de reconocimiento del test de realizaciones de habla con *stress* del experimento FFT-2b (FFT-Bandas Energía [20 bandas + 20 bandas delta <5 tramas>]).

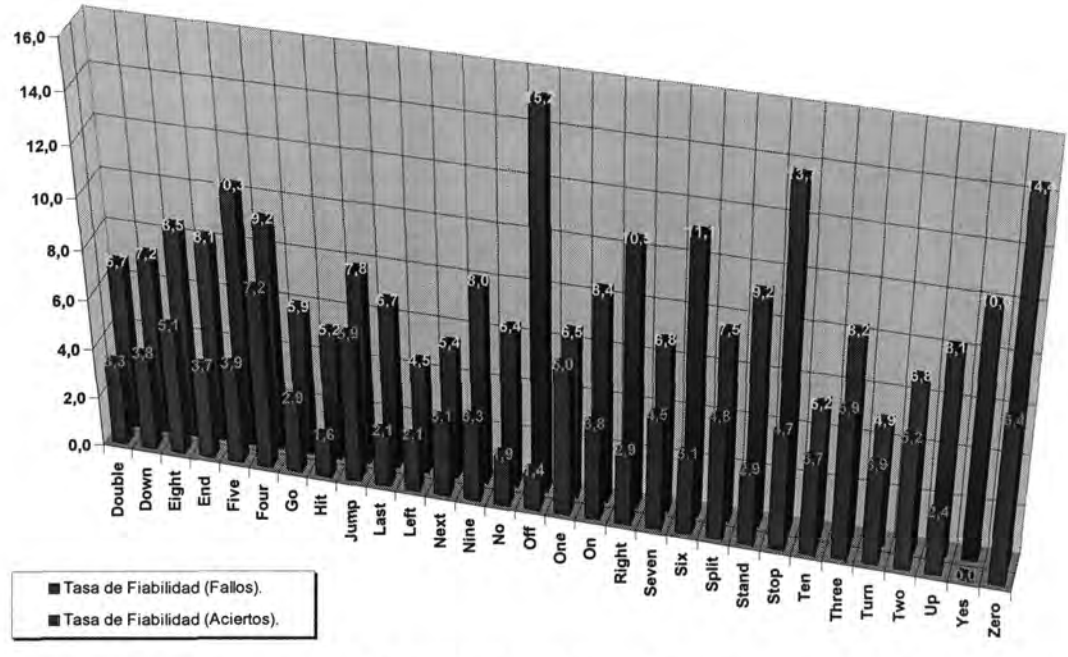

Figura B.18. Tasas de fiabilidad del test de realizaciones de habla con *stress* del experimento FFT-2b (FFT-Bandas Energía [20 bandas + 20 bandas delta <5 tramas>]).

	Double	Down	Eight	End	Five	Four	Go	Hit	Jump	Last	Left	Next	Nine	No	Off	One	On	Right	Seven	Six	Split	Stand	Stop	Ten	Three	Turn	Two	Up	Yes	Zero
Double	66%	8%	0%	0%	4%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	4%	0%	0%	8%	0%	0%	8%	0%	0%	0%	0%	0%	0%	0%	0%
Down	8%	62%	0%	0%	4%	0%	16%	0%	0%	0%	0%	0%	0%	4%	0%	0%	4%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Eight	0%	0%	50%	8%	0%	0%	0%	4%	0%	0%	0%	4%	0%	0%	0%	0%	0%	0%	0%	4%	20%	4%	0%	0%	4%	0%	0%	0%	0%	0%
End	0%	0%	4%	33%	0%	0%	0%	0%	0%	0%	0%	4%	0%	4%	0%	0%	4%	0%	4%	0%	0%	16%	0%	16%	4%	4%	4%	0%	0%	0%
Five	0%	0%	0%	0%	83%	4%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	12%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Four	0%	0%	0%	0%	0%	87%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	4%	0%	0%	0%	0%	4%	0%	0%	0%	4%	0%	0%	0%	0%
Go	33%	4%	0%	0%	0%	0%	41%	0%	0%	0%	0%	0%	0%	20%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Hit	0%	0%	4%	0%	0%	0%	0%	70%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	12%	8%	0%	0%	0%	4%	0%	0%	0%	0%	0%
Jump	0%	8%	0%	0%	0%	0%	4%	0%	83%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	4%	0%	0%	0%	0%	0%	0%	0%	0%
Last	0%	0%	0%	0%	0%	0%	0%	0%	0%	33%	62%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	4%	0%	0%
Left	0%	0%	4%	0%	0%	0%	0%	0%	0%	4%	70%	20%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Next	0%	0%	0%	0%	0%	0%	0%	4%	0%	0%	16%	75%	0%	0%	0%	0%	0%	0%	0%	4%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Nine	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	79%	4%	0%	4%	0%	0%	0%	0%	0%	12%	0%	0%	0%	0%	0%	0%	0%	0%
No	16%	4%	0%	0%	0%	0%	4%	0%	0%	0%	0%	0%	0%	75%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Off	4%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	87%	0%	4%	0%	0%	0%	0%	0%	4%	0%	0%	0%	0%	0%	0%	0%
One	4%	4%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	29%	8%	0%	45%	0%	0%	0%	0%	0%	8%	0%	0%	0%	0%	0%	0%	0%	0%
On	8%	0%	0%	0%	4%	0%	4%	0%	0%	0%	0%	0%	4%	0%	0%	8%	58%	0%	0%	0%	0%	12%	0%	0%	0%	0%	0%	0%	0%	0%
Right	0%	0%	0%	0%	4%	0%	0%	0%	0%	0%	4%	0%	0%	0%	0%	0%	0%	87%	4%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Seven	4%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	79%	0%	0%	4%	0%	4%	0%	4%	0%	0%	0%	4%
Six	0%	0%	0%	0%	0%	0%	0%	8%	0%	0%	0%	4%	0%	0%	0%	0%	0%	0%	0%	79%	8%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Split	0%	0%	4%	0%	0%	0%	0%	4%	0%	4%	0%	0%	0%	0%	0%	0%	0%	0%	0%	8%	79%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Stand	4%	0%	0%	4%	0%	0%	0%	0%	0%	4%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	83%	0%	0%	0%	4%	0%	0%	0%	0%
Stop	4%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	83%	0%	0%	0%	0%	12%	0%	0%
Ten	0%	0%	4%	16%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	4%	0%	62%	0%	4%	0%	0%	0%	8%
Three	0%	0%	0%	4%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	4%	4%	0%	0%	79%	0%	4%	0%	0%	4%
Turn	0%	12%	0%	0%	4%	0%	12%	0%	0%	0%	0%	0%	0%	0%	0%	4%	8%	0%	0%	0%	0%	8%	0%	4%	0%	45%	0%	0%	0%	0%
Two	12%	4%	0%	4%	0%	0%	4%	0%	4%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	8%	0%	16%	41%	0%	4%
Up	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	8%	0%	0%	0%	0%	0%	0%	0%	12%	0%	0%	0%	0%	79%	0%	0%
Yes	0%	0%	0%	0%	0%	0%	4%	0%	0%	0%	4%	20%	0%	0%	0%	0%	0%	0%	0%	0%	4%	0%	0%	0%	0%	0%	0%	0%	66%	0%
Zero	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	8%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	91%

Tabla B.9. Ejemplo de Tabla B.de confusión del test de realizaciones con stress del experimento FFT-2b (FFT-Bandas Energía [20 bandas + 20 bandas delta <5 tramas>]).

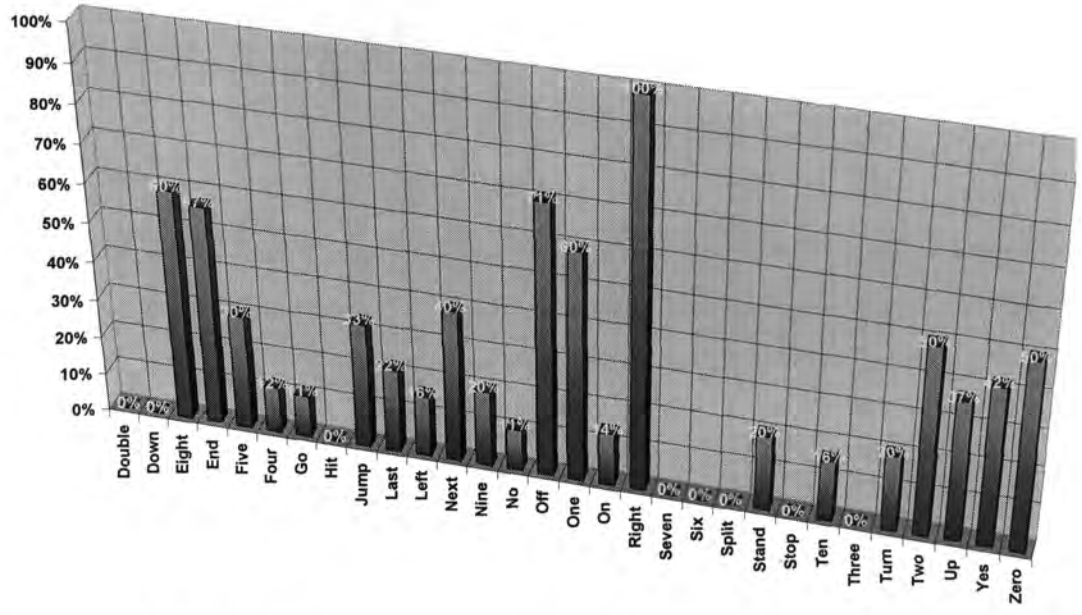

Figura B.19. Tasas de reconocimiento del test de realizaciones ruidosas del experimento FFT-8b (FFT-Cepstrum [10 coeficientes + 10 coeficientes delta <5 tramas>]).

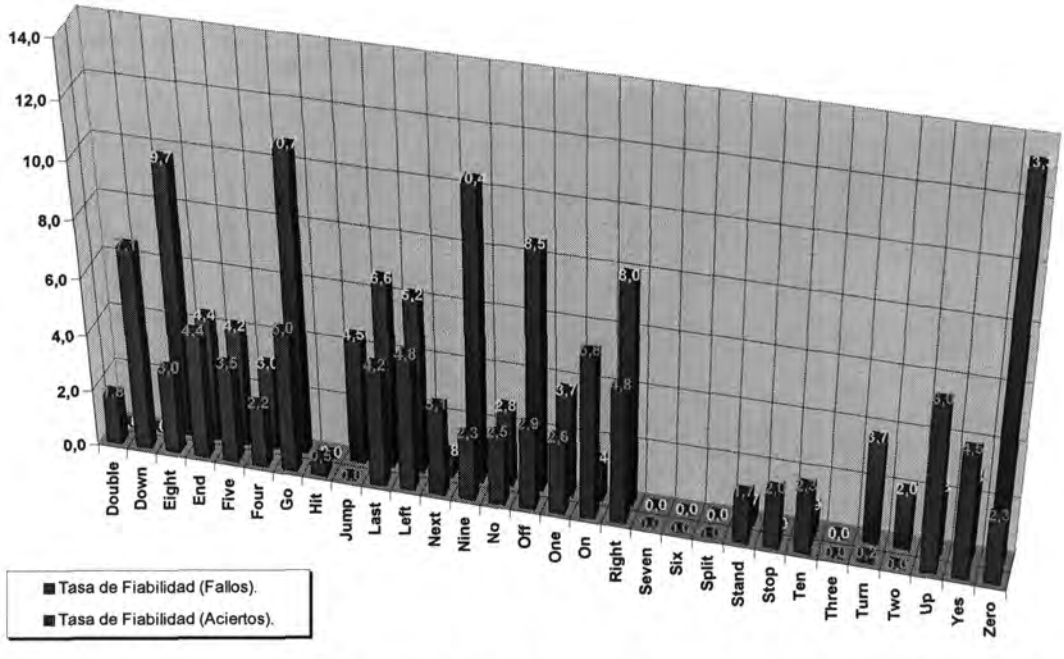

Figura B.20. Tasas de fiabilidad del test de realizaciones ruidosas del experimento FFT-8b (FFT-Cepstrum [10 coeficientes + 10 coeficientes delta <5 tramas>]).

	Double	Down	Eight	End	Five	Four	Go	Hit	Jump	Last	Left	Next	Nine	No	Off	One	On	Right	Seven	Six	Split	Stand	Stop	Ten	Three	Turn	Two	Up	Yes	Zero
Double	0%	12%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	37%	0%	12%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	37%	0%	0%
Down	0%	0%	0%	0%	12%	0%	0%	0%	0%	12%	0%	0%	0%	0%	0%	0%	0%	25%	0%	0%	0%	0%	0%	0%	0%	0%	0%	50%	0%	0%
Eight	0%	0%	60%	0%	0%	0%	0%	0%	0%	0%	20%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	20%	0%	0%	0%	0%	0%	0%	0%	0%
End	0%	0%	14%	57%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	14%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	14%	0%
Five	0%	0%	0%	0%	30%	0%	0%	0%	0%	0%	0%	0%	0%	0%	10%	0%	10%	50%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Four	0%	0%	0%	0%	25%	12%	0%	0%	0%	0%	0%	0%	0%	0%	37%	25%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Go	0%	0%	0%	0%	22%	11%	11%	0%	0%	11%	0%	0%	0%	0%	22%	0%	22%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Hit	0%	0%	100%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Jump	16%	0%	0%	0%	33%	0%	0%	0%	33%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	16%	0%	0%
Last	0%	0%	0%	0%	33%	0%	0%	0%	0%	22%	11%	0%	0%	0%	0%	0%	0%	22%	0%	0%	0%	0%	0%	11%	0%	0%	0%	0%	0%	0%
Left	0%	0%	0%	16%	0%	0%	0%	0%	0%	0%	16%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	16%	0%	50%	0%	0%	0%	0%	0%	0%
Next	0%	0%	0%	60%	0%	0%	0%	0%	0%	0%	0%	40%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Nine	0%	0%	0%	0%	30%	0%	0%	0%	0%	10%	0%	0%	20%	0%	0%	0%	0%	30%	0%	0%	0%	0%	0%	0%	0%	0%	0%	10%	0%	0%
No	0%	0%	0%	0%	11%	0%	44%	0%	0%	0%	0%	0%	0%	11%	0%	0%	11%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	11%	0%	11%
Off	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	71%	0%	14%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	14%	0%	0%
One	0%	0%	0%	0%	10%	0%	0%	0%	0%	10%	0%	0%	0%	0%	0%	60%	0%	20%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
On	0%	0%	0%	0%	14%	0%	0%	0%	0%	0%	0%	0%	0%	0%	42%	0%	14%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	28%	0%	0%
Right	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	100%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Seven	0%	0%	0%	14%	0%	0%	14%	14%	0%	0%	0%	0%	0%	0%	0%	0%	0%	14%	0%	0%	0%	0%	0%	14%	0%	14%	0%	0%	14%	0%
Six	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	100%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Split	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	50%	50%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Stand	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	20%	0%	0%	20%	0%	20%	0%	0%	0%	20%	0%	20%	0%	0%	0%	0%	0%	0%
Stop	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	20%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	80%	0%	0%
Ten	0%	0%	0%	16%	0%	0%	0%	0%	0%	16%	0%	16%	0%	0%	0%	0%	0%	0%	0%	0%	0%	16%	0%	16%	0%	0%	0%	0%	0%	16%
Three	0%	0%	0%	66%	0%	0%	0%	33%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Turn	0%	0%	0%	0%	40%	0%	0%	0%	0%	10%	0%	0%	10%	0%	0%	0%	0%	10%	0%	0%	0%	0%	0%	0%	0%	20%	0%	10%	0%	0%
Two	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	50%	50%	0%	0%
Up	0%	0%	0%	0%	25%	0%	0%	0%	0%	0%	0%	0%	0%	0%	12%	0%	0%	25%	0%	0%	0%	0%	0%	0%	0%	0%	0%	37%	0%	0%
Yes	0%	0%	28%	0%	0%	0%	0%	0%	0%	0%	14%	14%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	42%	0%
Zero	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	25%	0%	0%	0%	0%	0%	25%	50%

Tabla B.10. Ejemplo de Tabla B.de confusión del test de realizaciones ruidosas del experimento FFT-8b (FFT-Cepstrum [10 coeficientes + 10 coeficientes delta <5 tramas>]).

B.3. PARAMETRIZADORES BASADOS EN PLP

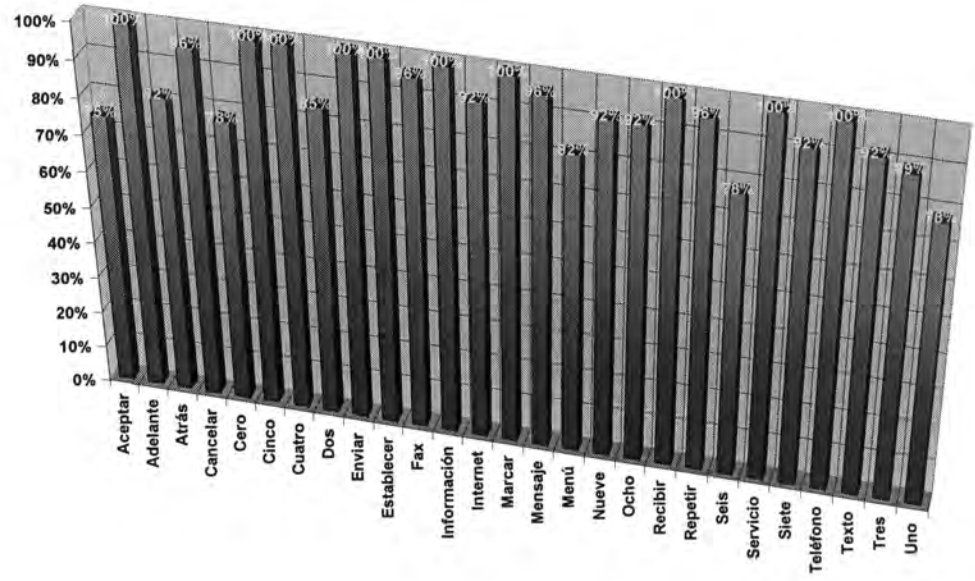

Figura B.21. Tasas de reconocimiento del test de realizaciones con entonación enunciativa del experimento PLP-5a (PLP sin normalización de energía en LPC-Cepstrum [10 coeficientes + 10 coeficientes delta <5 tramas>]).

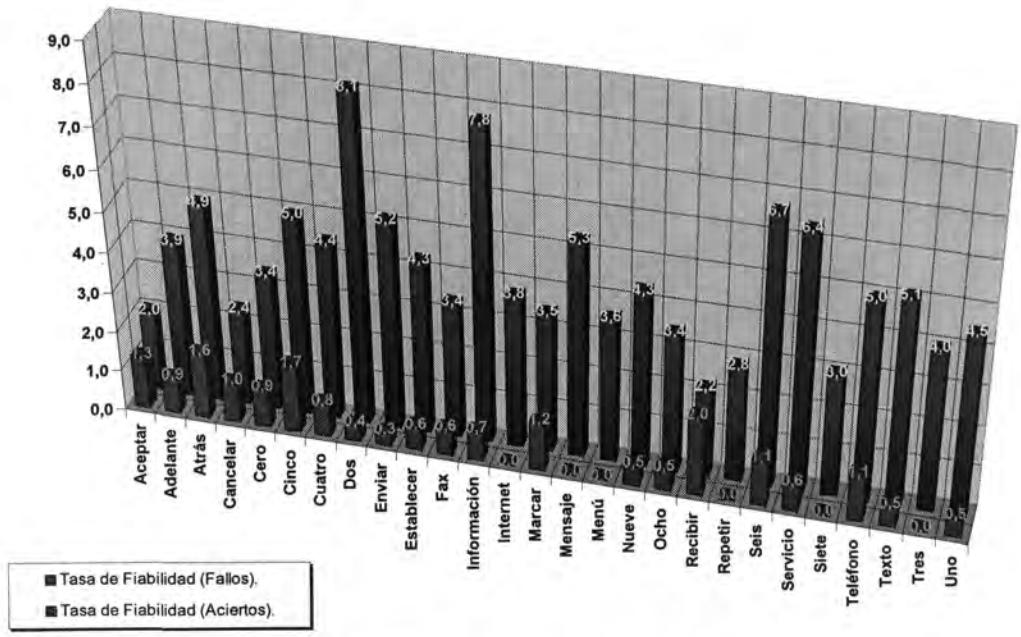

Figura B.22. Tasas de fiabilidad del test de realizaciones con entonación enunciativa del experimento PLP-5a (PLP sin normalización de energía en LPC-Cepstrum [10 coeficientes + 10 coeficientes delta <5 tramas>]).

	Aceptar	Adelante	Atrás	Cancelar	Cero	Cinco	Cuatro	Dos	Enviar	Establecer	Fax	Información	Internet	Marcar	Mensaje	Menú	Nueve	Ocho	Recibir	Repetir	Seis	Servicio	Siete	Teléfono	Texto	Tres	Uno
Aceptar	75%	3%	0%	10%	0%	0%	0%	0%	0%	0%	0%	0%	0%	10%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Adelante	0%	100%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Atrás	3%	0%	82%	3%	0%	0%	0%	3%	0%	0%	7%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Cancelar	3%	0%	0%	96%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Cero	0%	0%	0%	0%	78%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	21%	0%	0%	0%
Cinco	0%	0%	0%	0%	0%	100%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Cuatro	0%	0%	0%	0%	0%	0%	100%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Dos	0%	0%	3%	0%	0%	0%	0%	85%	0%	0%	0%	3%	0%	0%	0%	0%	0%	7%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Enviar	0%	0%	0%	0%	0%	0%	0%	0%	100%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Establecer	0%	0%	0%	0%	0%	0%	0%	0%	0%	100%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Fax	0%	0%	3%	0%	0%	0%	0%	0%	0%	0%	96%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Información	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	100%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Internet	0%	0%	0%	3%	0%	0%	0%	0%	3%	0%	0%	0%	92%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Marcar	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	100%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Mensaje	0%	0%	0%	0%	0%	0%	0%	0%	0%	3%	0%	0%	0%	0%	96%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Menú	0%	0%	0%	0%	0%	3%	3%	0%	0%	0%	0%	3%	0%	0%	0%	82%	3%	0%	0%	0%	0%	0%	0%	0%	0%	0%	3%
Nueve	0%	7%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	92%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Ocho	0%	0%	0%	0%	0%	0%	3%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	92%	0%	0%	0%	0%	0%	3%	0%	0%	0%
Recibir	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	100%	0%	0%	0%	0%	0%	0%	0%	0%
Repetir	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	3%	96%	0%	0%	0%	0%	0%	0%	0%
Seis	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	7%	0%	78%	7%	0%	7%	0%	0%	0%
Servicio	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	100%	0%	0%	0%	0%	0%
Siete	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	3%	3%	0%	0%	92%	0%	0%	0%	0%
Teléfono	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	100%	0%	0%	0%
Texto	0%	0%	0%	0%	3%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	3%	0%	0%	92%	0%	0%
Tres	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	7%	0%	0%	0%	3%	89%	0%
Uno	0%	0%	0%	0%	0%	7%	7%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	3%	0%	0%	0%	0%	0%	3%	0%	0%	78%

Tabla B.11. Ejemplo de Tabla B.de confusión del test de realizaciones con entonación enunciativa del experimento PLP-5a (PLP sin normalización de energía en LPC-Cepstrum [10 coeficientes + 10 coeficientes delta <5 tramas>]).

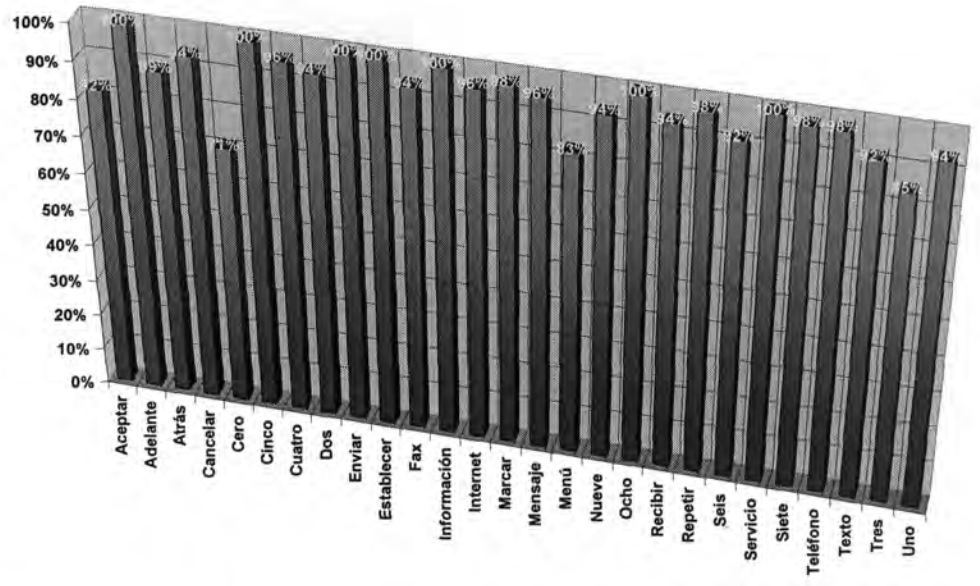

Figura B.23. Tasas de reconocimiento del test de realizaciones con entonación exclamativa/interrogativa del experimento PLP-7a (PLP sin tratamiento perceptual [10 coeficientes + 10 coeficientes delta <5 tramas>]).

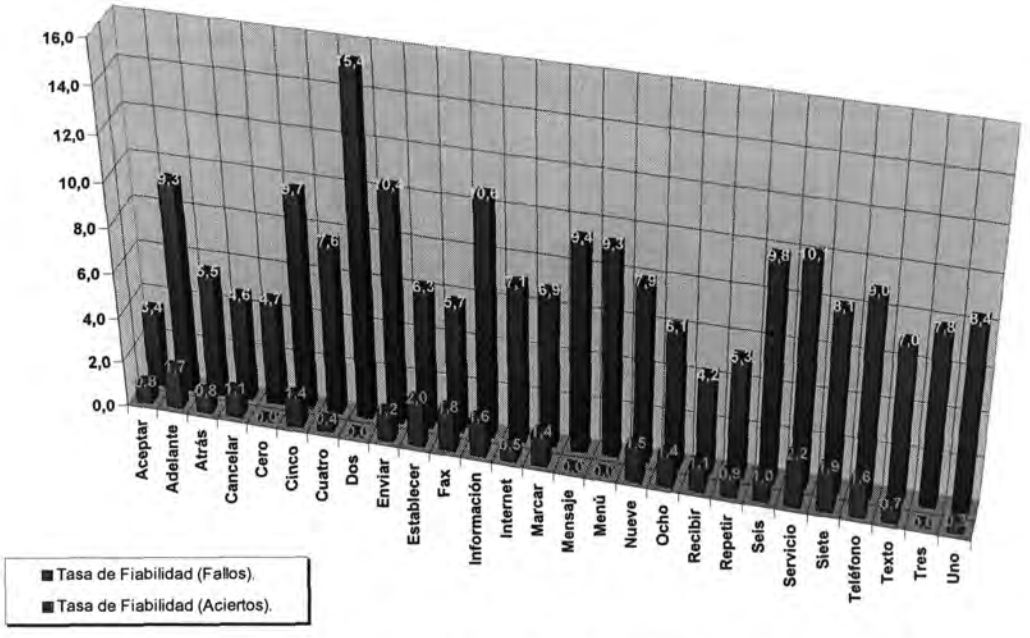

Figura B.24. Tasas de fiabilidad del test de realizaciones con entonación exclamativa/interrogativa del experimento PLP-7a (PLP sin tratamiento perceptual [10 coeficientes + 10 coeficientes delta <5 tramas>]).

	Aceptar	Adelante	Atrás	Cancelar	Cero	Cinco	Cuatro	Dos	Enviar	Establecer	Fax	Información	Internet	Marcar	Mensaje	Menú	Nueve	Ocho	Recibir	Repetir	Seis	Servicio	Siete	Teléfono	Texto	Tres	Uno
Aceptar	82%	0%	7%	8%	0%	0%	0%	0%	0%	0%	0%	0%	0%	1%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Adelante	0%	100%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Atrás	3%	0%	89%	5%	0%	0%	0%	0%	0%	0%	1%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Cancelar	3%	0%	0%	94%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	1%
Cero	0%	0%	0%	0%	71%	5%	1%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	1%	0%	10%	8%	0%	0%
Cinco	0%	0%	0%	0%	0%	100%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Cuatro	0%	0%	0%	0%	0%	0%	96%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	3%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Dos	0%	0%	1%	0%	0%	0%	0%	94%	0%	0%	0%	1%	0%	0%	0%	0%	0%	1%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Enviar	0%	0%	0%	0%	0%	0%	0%	0%	100%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Establecer	0%	0%	0%	0%	0%	0%	0%	0%	0%	100%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Fax	3%	0%	1%	0%	0%	0%	0%	0%	0%	0%	94%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Información	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	100%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Internet	0%	1%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	96%	1%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Marcar	1%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	98%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Mensaje	0%	0%	0%	0%	0%	0%	0%	0%	0%	1%	0%	0%	1%	0%	96%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Menú	0%	0%	0%	0%	0%	1%	0%	0%	3%	0%	0%	1%	5%	0%	0%	83%	1%	0%	0%	0%	0%	0%	0%	1%	0%	0%	0%
Nueve	0%	3%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	1%	0%	0%	0%	94%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Ocho	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	100%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Recibir	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	94%	5%	0%	0%	0%	0%	0%	0%	0%
Repetir	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	1%	0%	0%	0%	0%	0%	0%	98%	0%	0%	0%	0%	0%	0%	0%
Seis	0%	0%	0%	0%	0%	0%	0%	0%	0%	1%	0%	0%	0%	0%	0%	0%	0%	0%	3%	0%	92%	1%	0%	0%	0%	0%	0%
Servicio	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	100%	0%	0%	0%	0%	0%
Siete	0%	1%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	98%	0%	0%	0%	0%
Teléfono	0%	0%	0%	0%	0%	1%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	98%	0%	0%	0%
Texto	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	1%	5%	92%	0%	0%
Tres	0%	0%	0%	0%	0%	0%	0%	0%	0%	7%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	5%	0%	0%	1%	0%	85%	0%
Uno	0%	0%	0%	0%	0%	1%	0%	0%	0%	0%	0%	3%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	94%

Tabla B.12. Ejemplo de Tabla B.de confusión del test de realizaciones con entonaciones exclamativa e interrogativa del experimento PLP-7a (PLP sin tratamiento perceptual [10 coeficientes + 10 coeficientes delta <5 tramas>]).

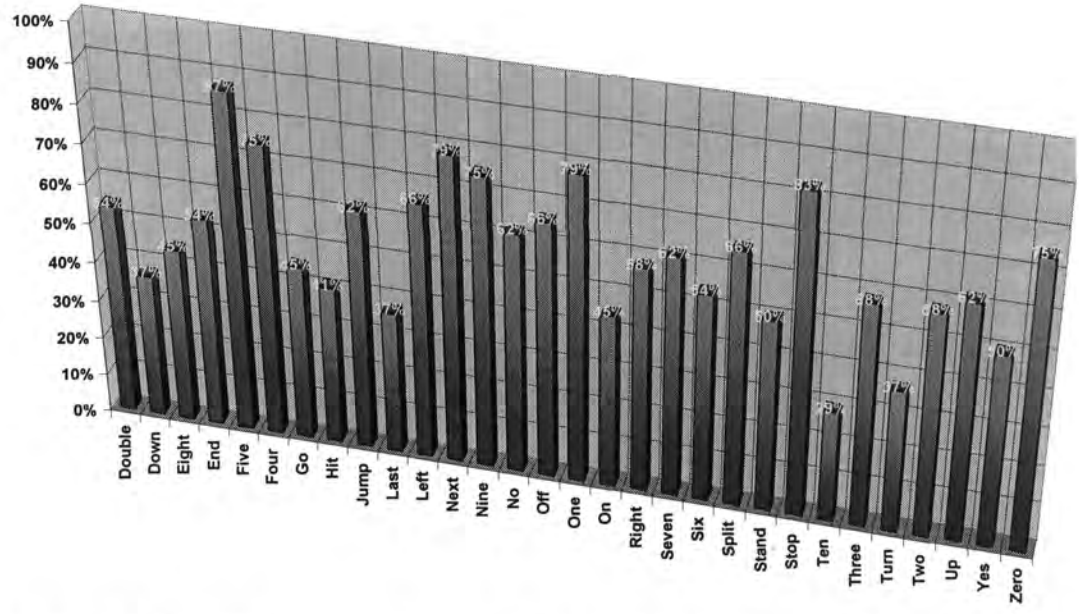

Figura B.25. Tasas de reconocimiento del test de realizaciones silenciosas del experimento PLP-2b (PLP [10 coeficientes + 10 coeficientes delta <5 tramas>]).

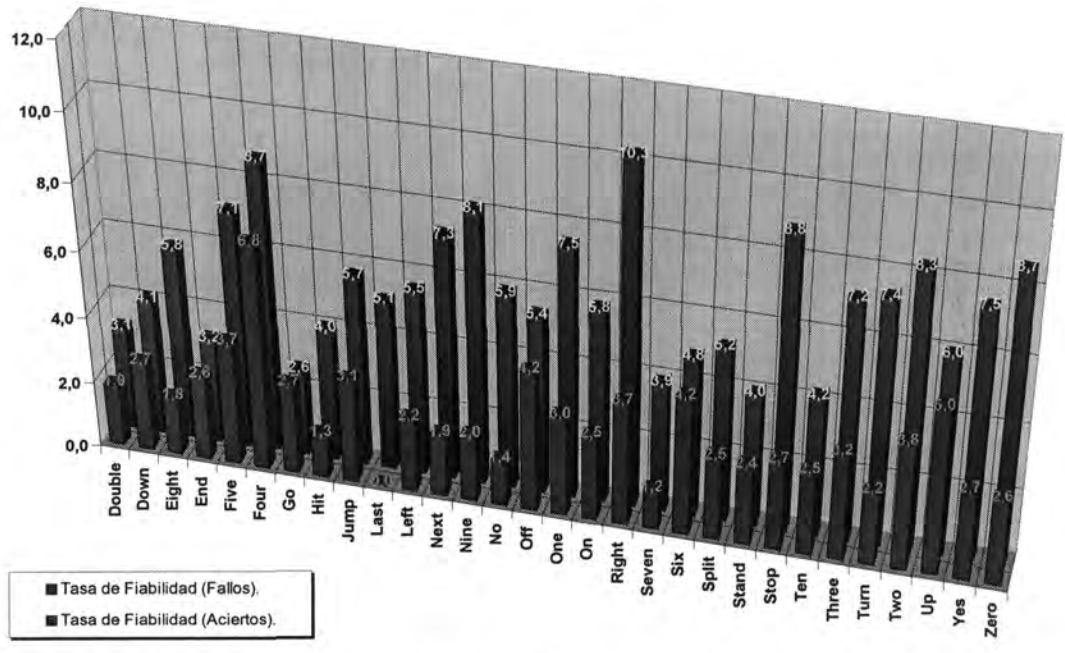

Figura B.26. Tasas de fiabilidad del test de realizaciones silenciosas del experimento PLP-2b (PLP [10 coeficientes + 10 coeficientes delta <5 tramas>]).

	Double	Down	Eight	End	Five	Four	Go	Hit	Jump	Last	Left	Next	Nine	No	Off	One	On	Right	Seven	Six	Split	Stand	Stop	Ten	Three	Turn	Two	Up	Yes	Zero
Double	54%	0%	0%	0%	0%	0%	16%	0%	0%	0%	8%	0%	4%	8%	0%	0%	0%	0%	4%	0%	4%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Down	12%	37%	0%	0%	0%	0%	20%	0%	4%	0%	0%	0%	4%	12%	0%	0%	0%	0%	4%	0%	4%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Eight	0%	0%	45%	20%	0%	0%	0%	4%	0%	0%	0%	8%	0%	0%	0%	4%	0%	0%	0%	0%	8%	4%	0%	0%	0%	0%	4%	0%	0%	0%
End	0%	4%	0%	54%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	8%	0%	0%	0%	4%	0%	8%	0%	4%	4%	0%	4%	0%	0%	8%
Five	0%	4%	0%	0%	87%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	4%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	4%
Four	4%	0%	0%	0%	0%	75%	8%	0%	0%	0%	0%	0%	0%	0%	4%	0%	0%	0%	0%	0%	0%	4%	4%	0%	0%	0%	0%	0%	0%	0%
Go	16%	4%	0%	0%	0%	8%	45%	0%	0%	0%	0%	0%	0%	12%	4%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	8%
Hit	0%	4%	4%	0%	4%	0%	0%	41%	0%	0%	4%	0%	0%	0%	0%	0%	0%	4%	4%	4%	8%	0%	0%	0%	20%	0%	0%	0%	0%	0%
Jump	4%	20%	0%	8%	0%	0%	0%	0%	62%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	4%	0%	0%	0%	0%	0%	0%	0%
Last	0%	0%	0%	0%	0%	0%	4%	0%	0%	37%	33%	0%	0%	0%	0%	0%	4%	0%	0%	0%	0%	8%	4%	0%	0%	0%	0%	0%	4%	4%
Left	0%	0%	4%	0%	0%	0%	0%	0%	0%	0%	66%	8%	4%	0%	0%	0%	0%	0%	4%	0%	0%	4%	0%	0%	0%	0%	0%	0%	4%	4%
Next	4%	0%	0%	4%	0%	0%	0%	0%	0%	0%	0%	79%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	12%	0%
Nine	4%	12%	0%	4%	0%	0%	0%	0%	0%	0%	0%	0%	75%	0%	0%	4%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
No	4%	12%	0%	0%	0%	0%	16%	0%	0%	0%	0%	0%	0%	62%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	4%
Off	0%	0%	0%	0%	0%	8%	0%	0%	0%	0%	4%	0%	0%	0%	66%	0%	4%	0%	0%	0%	0%	8%	0%	0%	0%	0%	8%	0%	0%	0%
One	4%	0%	0%	0%	0%	0%	4%	0%	0%	0%	0%	0%	4%	0%	0%	79%	8%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
On	4%	0%	0%	0%	0%	4%	0%	0%	4%	0%	0%	0%	0%	0%	0%	37%	45%	0%	0%	0%	0%	0%	0%	0%	0%	4%	0%	0%	0%	0%
Right	4%	8%	0%	4%	16%	0%	0%	0%	0%	0%	0%	0%	4%	0%	0%	0%	0%	58%	0%	0%	0%	0%	0%	4%	0%	0%	0%	0%	0%	0%
Seven	0%	0%	0%	4%	0%	0%	0%	0%	0%	0%	0%	0%	4%	0%	0%	0%	0%	0%	62%	0%	0%	4%	0%	4%	0%	4%	4%	0%	0%	12%
Six	0%	0%	0%	0%	0%	0%	0%	8%	0%	0%	0%	8%	0%	0%	0%	0%	0%	0%	0%	54%	25%	4%	0%	0%	0%	0%	0%	0%	0%	0%
Split	0%	0%	0%	0%	0%	0%	0%	4%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	66%	8%	0%	0%	16%	0%	0%	0%	0%	4%
Stand	0%	4%	0%	0%	4%	0%	0%	0%	4%	0%	0%	0%	0%	0%	0%	0%	0%	0%	20%	0%	0%	50%	0%	0%	0%	4%	0%	0%	0%	12%
Stop	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	4%	83%	0%	0%	0%	4%	4%	0%	4%
Ten	0%	0%	4%	16%	0%	0%	4%	0%	0%	0%	0%	0%	8%	0%	0%	0%	0%	4%	4%	0%	0%	8%	0%	29%	0%	12%	4%	0%	0%	4%
Three	0%	4%	0%	4%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	4%	0%	25%	4%	0%	0%	58%	0%	0%	0%	0%	0%
Turn	4%	12%	0%	4%	0%	0%	0%	0%	4%	0%	0%	0%	4%	0%	0%	0%	4%	0%	8%	0%	0%	4%	0%	12%	0%	37%	0%	0%	0%	4%
Two	4%	4%	0%	0%	4%	0%	8%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	4%	0%	8%	58%	0%	0%	8%
Up	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	4%	0%	0%	0%	8%	0%	0%	0%	0%	0%	0%	0%	25%	0%	0%	0%	0%	62%	0%	0%
Yes	0%	0%	0%	0%	0%	0%	0%	4%	0%	0%	4%	12%	0%	4%	0%	0%	0%	0%	0%	0%	8%	4%	0%	0%	0%	0%	0%	0%	50%	12%
Zero	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	12%	0%	0%	0%	4%	0%	0%	0%	8%	0%	0%	75%

Tabla B.13. Ejemplo de Tabla B.de confusión del test de realizaciones silenciosas del experimento PLP-2b (PLP [10 coeficientes + 10 coeficientes delta <5 tramas>]).

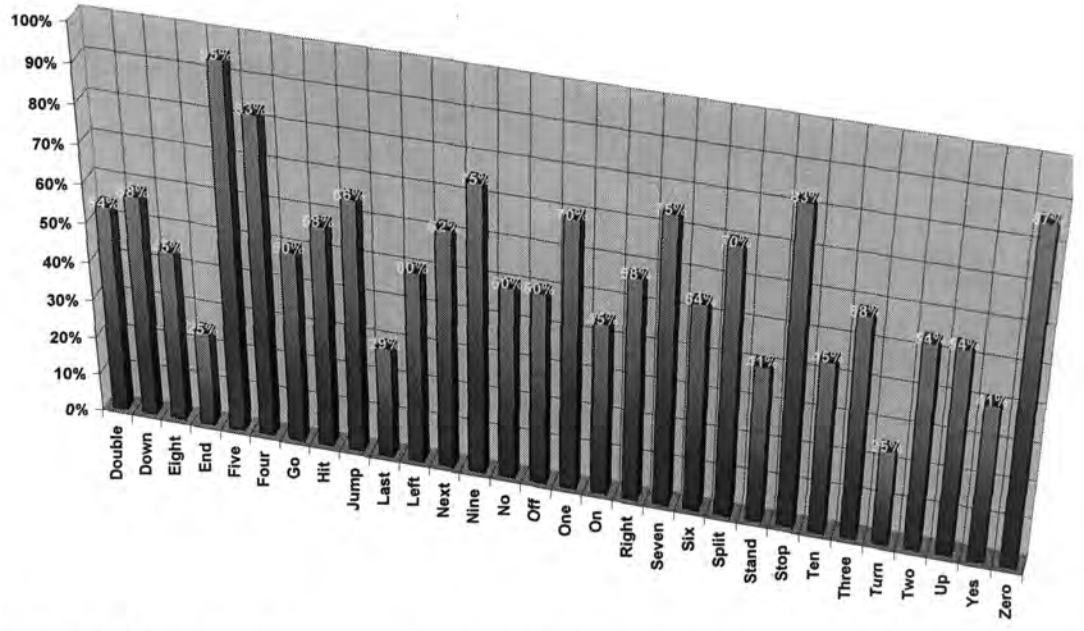

Figura B.27. Tasas de reconocimiento del test de realizaciones de habla con *stress* del experimento PLP-3b (PLP [10 coeficientes + 10 coeficientes delta <5 tramas> + 10 coeficientes delta-delta]).

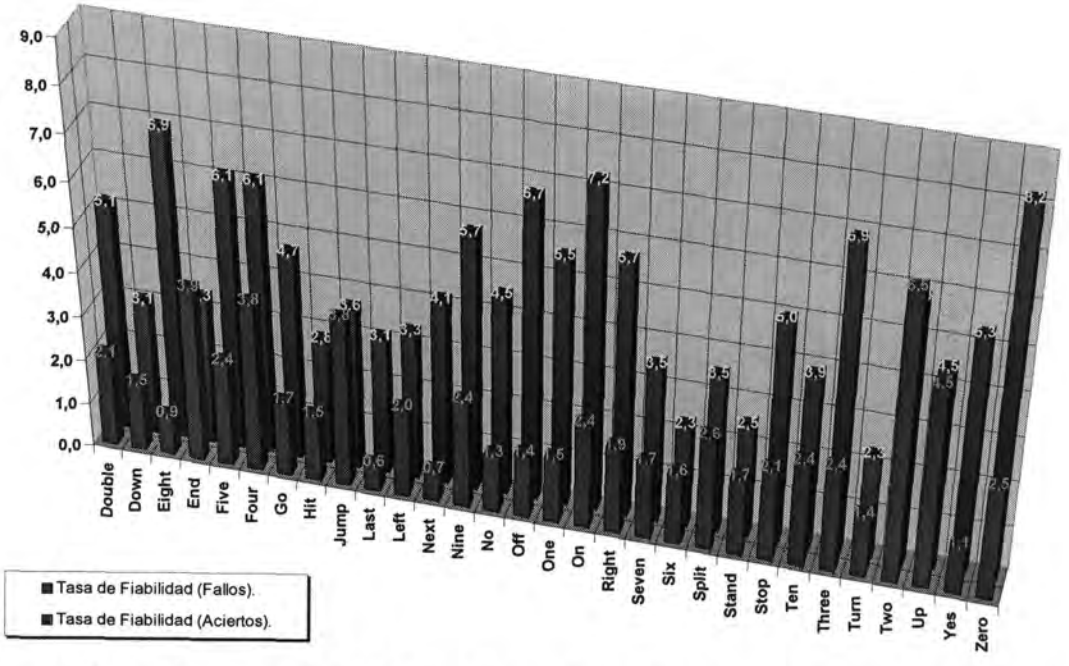

Figura B.28. Tasas de fiabilidad del test de realizaciones de habla con *stress* del experimento PLP-3b (PLP [10 coeficientes + 10 coeficientes delta <5 tramas> + 10 coeficientes delta-delta]).

	Double	Down	Eight	End	Five	Four	Go	Hit	Jump	Last	Left	Next	Nine	No	Off	One	On	Right	Seven	Six	Split	Stand	Stop	Ten	Three	Turn	Two	Up	Yes	Zero
Double	54%	16%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	4%	0%	0%	0%	0%	0%	4%	0%	0%	8%	4%	0%	0%	0%	0%	0%	0%	8%
Down	12%	58%	0%	0%	0%	0%	4%	0%	4%	0%	0%	0%	4%	4%	0%	8%	0%	0%	0%	0%	0%	0%	0%	0%	0%	4%	0%	0%	0%	0%
Eight	0%	0%	45%	0%	4%	0%	0%	4%	0%	0%	0%	8%	0%	0%	0%	0%	0%	12%	4%	4%	12%	0%	0%	0%	4%	0%	0%	0%	0%	0%
End	0%	0%	4%	25%	4%	0%	0%	0%	4%	0%	0%	4%	0%	0%	0%	0%	0%	0%	29%	0%	4%	8%	0%	12%	4%	0%	0%	0%	0%	0%
Five	0%	0%	0%	0%	95%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	4%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Four	4%	0%	0%	0%	0%	83%	4%	0%	0%	0%	0%	0%	0%	0%	4%	4%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Go	33%	0%	0%	0%	0%	0%	50%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	16%
Hit	0%	0%	0%	0%	0%	0%	0%	58%	0%	0%	4%	0%	0%	0%	0%	0%	0%	0%	4%	8%	16%	0%	0%	0%	8%	0%	0%	0%	0%	0%
Jump	8%	12%	0%	0%	0%	0%	0%	0%	66%	0%	0%	0%	0%	0%	0%	0%	0%	0%	8%	0%	0%	4%	0%	0%	0%	0%	0%	0%	0%	0%
Last	0%	0%	0%	0%	4%	0%	0%	0%	4%	29%	25%	0%	0%	0%	0%	4%	0%	8%	4%	0%	0%	4%	4%	0%	0%	4%	0%	0%	8%	0%
Left	0%	0%	4%	0%	4%	0%	0%	0%	0%	4%	50%	0%	4%	0%	0%	0%	0%	0%	8%	0%	4%	4%	0%	0%	0%	4%	0%	0%	8%	4%
Next	0%	0%	0%	0%	4%	0%	0%	0%	0%	0%	12%	62%	4%	0%	0%	0%	0%	0%	4%	0%	8%	0%	0%	0%	0%	0%	0%	0%	4%	0%
Nine	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	75%	0%	0%	8%	0%	0%	4%	0%	0%	12%	0%	0%	0%	0%	0%	0%	0%	0%
No	25%	4%	0%	0%	0%	4%	4%	0%	0%	0%	0%	0%	0%	50%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	12%
Off	4%	0%	0%	0%	0%	16%	0%	0%	0%	0%	0%	0%	0%	0%	50%	0%	4%	0%	0%	0%	0%	0%	20%	0%	0%	0%	0%	4%	0%	0%
One	8%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	8%	0%	0%	70%	0%	0%	8%	0%	0%	4%	0%	0%	0%	0%	0%	0%	0%	0%
On	20%	8%	0%	0%	0%	12%	0%	0%	0%	0%	0%	0%	0%	0%	0%	8%	45%	0%	4%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Right	0%	0%	0%	0%	8%	0%	0%	0%	0%	0%	0%	4%	16%	0%	0%	4%	0%	58%	0%	0%	0%	0%	0%	0%	8%	0%	0%	0%	0%	0%
Seven	0%	0%	0%	0%	8%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	75%	0%	0%	12%	0%	0%	0%	0%	0%	0%	0%	4%
Six	0%	0%	0%	0%	0%	0%	0%	8%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	54%	37%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Split	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	70%	4%	0%	0%	20%	0%	0%	0%	0%	4%
Stand	8%	8%	0%	0%	4%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	33%	0%	0%	41%	0%	0%	0%	0%	0%	0%	0%	4%
Stop	0%	0%	0%	0%	0%	0%	4%	0%	4%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	83%	0%	0%	0%	0%	4%	0%	4%
Ten	0%	0%	0%	12%	0%	0%	0%	0%	0%	0%	0%	0%	8%	0%	0%	0%	0%	0%	12%	0%	4%	0%	0%	45%	0%	4%	0%	0%	4%	8%
Three	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	37%	0%	0%	0%	58%	0%	4%	0%	0%	0%
Turn	8%	12%	0%	4%	0%	0%	0%	0%	0%	0%	0%	0%	0%	4%	0%	12%	8%	0%	12%	4%	0%	0%	0%	0%	0%	25%	4%	0%	0%	4%
Two	4%	0%	0%	0%	0%	4%	4%	0%	0%	4%	0%	0%	0%	0%	0%	0%	0%	0%	4%	4%	4%	4%	4%	4%	0%	0%	54%	0%	0%	4%
Up	0%	4%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	4%	0%	0%	4%	0%	0%	0%	0%	33%	0%	0%	0%	0%	54%	0%	0%
Yes	0%	0%	12%	0%	0%	0%	0%	8%	0%	0%	4%	8%	0%	0%	0%	0%	0%	0%	0%	0%	4%	8%	0%	8%	0%	0%	0%	0%	41%	4%
Zero	8%	4%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	87%

Tabla B.14. Ejemplo de Tabla B.de confusión del test de realizaciones con stress del experimento PLP-3b (PLP [10 coeficientes + 10 coeficientes delta <5 tramas> + 10 coeficientes delta-delta]).

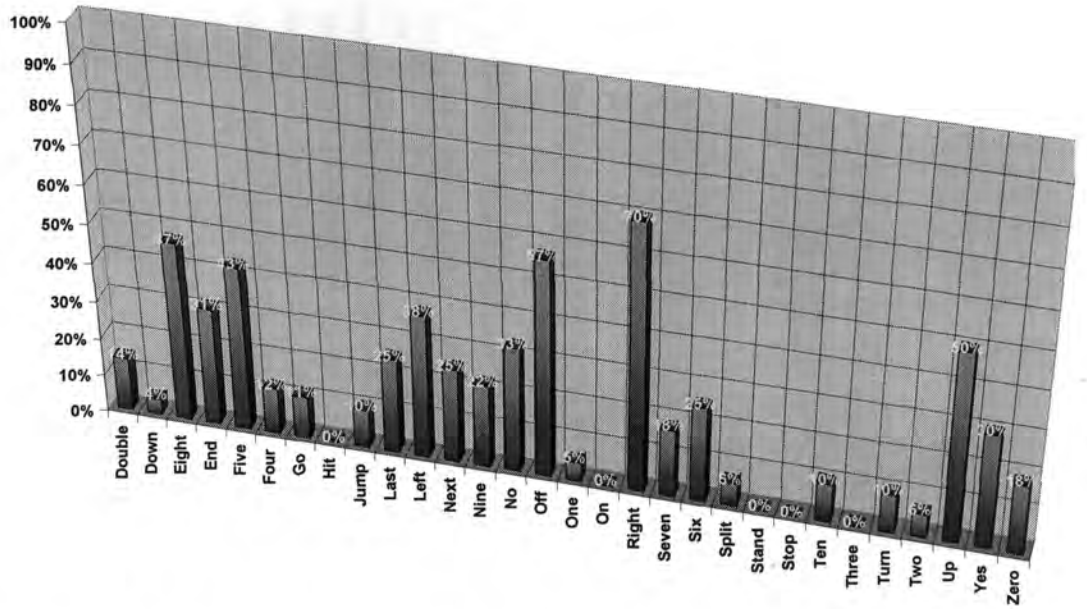

Figura B.29. Tasas de reconocimiento del test de realizaciones ruidosas del experimento PLP-3b (PLP [10 coeficientes + 10 coeficientes delta <5 tramas> + 10 coeficientes delta-delta]).

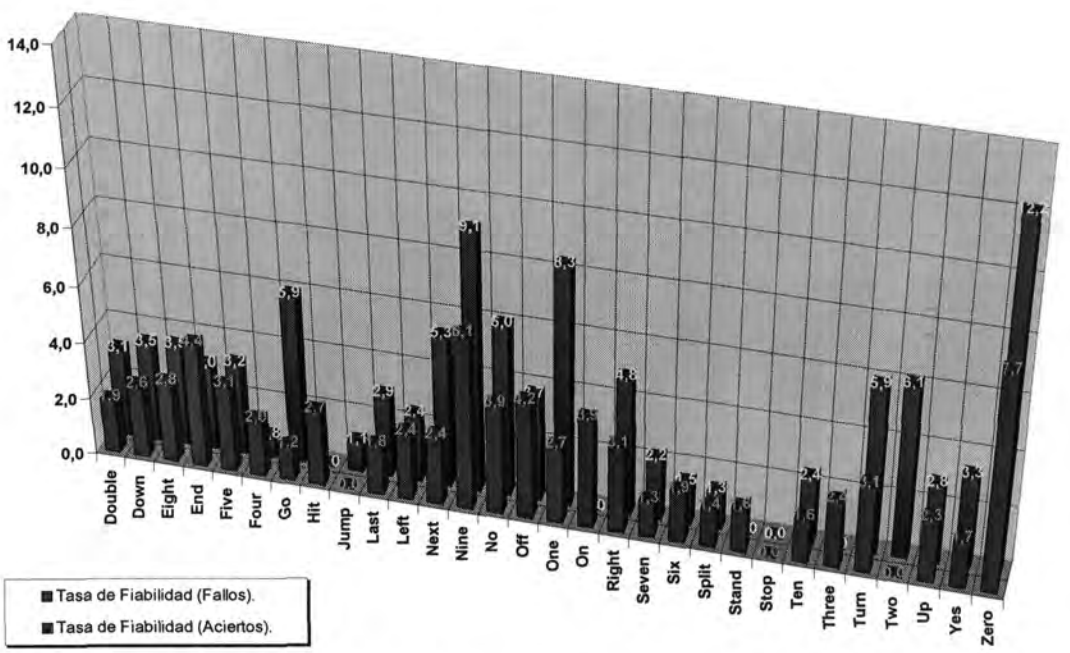

Figura B.30. Tasas de fiabilidad del test de realizaciones ruidosas del experimento PLP-3b (PLP [10 coeficientes + 10 coeficientes delta <5 tramas> + 10 coeficientes delta-delta]).

	Double	Down	Eight	End	Five	Four	Go	Hit	Jump	Last	Left	Next	Nine	No	Off	One	On	Right	Seven	Six	Split	Stand	Stop	Ten	Three	Turn	Two	Up	Yes	Zero
Double	14%	0%	0%	0%	9%	9%	0%	0%	0%	0%	4%	4%	0%	14%	0%	4%	0%	9%	0%	0%	0%	4%	0%	0%	0%	4%	0%	4%	14%	0%
Down	4%	4%	0%	0%	13%	0%	4%	0%	0%	4%	4%	4%	4%	18%	0%	0%	0%	18%	0%	0%	0%	9%	0%	0%	0%	0%	0%	9%	0%	0%
Eight	0%	0%	47%	11%	0%	0%	0%	0%	0%	0%	0%	17%	0%	0%	0%	0%	0%	5%	5%	0%	0%	5%	0%	5%	0%	0%	0%	0%	0%	0%
End	0%	0%	0%	31%	18%	0%	0%	0%	0%	0%	6%	0%	0%	0%	0%	0%	0%	0%	6%	12%	0%	0%	0%	6%	0%	6%	0%	0%	12%	0%
Five	0%	0%	0%	0%	43%	0%	0%	0%	0%	4%	4%	4%	0%	0%	0%	4%	0%	21%	13%	0%	0%	0%	0%	0%	0%	0%	0%	4%	0%	0%
Four	0%	0%	12%	18%	6%	12%	0%	0%	0%	0%	6%	0%	0%	6%	12%	12%	0%	0%	0%	0%	0%	0%	0%	6%	0%	6%	0%	0%	0%	0%
Go	5%	0%	0%	11%	5%	0%	11%	0%	0%	0%	11%	5%	0%	22%	0%	5%	16%	5%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Hit	0%	0%	16%	0%	0%	0%	0%	0%	0%	0%	25%	25%	0%	0%	0%	0%	0%	0%	0%	8%	0%	8%	0%	0%	0%	0%	0%	0%	16%	0%
Jump	10%	0%	0%	0%	5%	0%	5%	0%	10%	10%	0%	5%	10%	10%	0%	0%	0%	0%	10%	0%	0%	0%	0%	0%	0%	5%	0%	15%	0%	0%
Last	5%	0%	0%	5%	15%	0%	0%	0%	0%	25%	0%	15%	10%	0%	0%	5%	0%	10%	0%	0%	0%	5%	0%	0%	0%	0%	0%	5%	0%	0%
Left	0%	0%	4%	0%	0%	0%	0%	0%	0%	4%	38%	23%	0%	0%	0%	0%	0%	0%	0%	4%	0%	9%	0%	4%	0%	0%	0%	0%	9%	0%
Next	0%	0%	0%	12%	0%	0%	0%	6%	0%	0%	18%	25%	0%	0%	0%	0%	0%	6%	0%	0%	0%	25%	0%	0%	0%	0%	0%	6%	0%	0%
Nine	0%	0%	0%	5%	11%	0%	0%	0%	0%	5%	11%	0%	22%	5%	0%	0%	0%	22%	0%	0%	0%	0%	0%	0%	0%	5%	0%	0%	11%	0%
No	22%	0%	0%	0%	5%	0%	5%	0%	0%	5%	0%	5%	0%	33%	0%	5%	0%	5%	0%	0%	0%	0%	0%	0%	0%	0%	0%	5%	0%	5%
Off	0%	0%	0%	0%	0%	7%	0%	0%	0%	0%	14%	7%	0%	0%	57%	7%	7%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
One	0%	0%	0%	0%	10%	0%	0%	0%	0%	5%	0%	0%	5%	0%	10%	5%	0%	47%	0%	0%	0%	0%	0%	5%	0%	0%	0%	10%	0%	0%
On	0%	0%	5%	17%	23%	0%	0%	0%	0%	0%	0%	0%	5%	0%	5%	29%	0%	11%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Right	0%	0%	0%	5%	15%	0%	0%	0%	0%	0%	5%	5%	0%	0%	0%	0%	0%	70%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Seven	0%	0%	0%	12%	0%	0%	0%	0%	0%	12%	18%	12%	0%	0%	0%	0%	0%	0%	18%	0%	0%	0%	0%	12%	0%	6%	0%	0%	8%	0%
Six	0%	0%	25%	0%	0%	0%	0%	0%	0%	0%	12%	25%	0%	0%	0%	0%	0%	0%	0%	25%	0%	0%	0%	12%	0%	0%	0%	0%	0%	0%
Split	0%	0%	13%	0%	6%	0%	0%	0%	0%	0%	6%	0%	0%	0%	0%	0%	0%	0%	6%	6%	6%	13%	0%	6%	13%	0%	0%	0%	20%	0%
Stand	0%	0%	0%	15%	36%	0%	0%	0%	0%	5%	0%	0%	5%	5%	0%	0%	0%	15%	5%	0%	0%	0%	0%	0%	0%	5%	0%	5%	0%	0%
Stop	6%	6%	0%	0%	12%	0%	0%	0%	0%	0%	16%	0%	0%	0%	12%	6%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	37%	0%	0%
Ten	0%	0%	5%	5%	5%	0%	0%	5%	0%	0%	21%	15%	0%	0%	0%	0%	0%	0%	0%	0%	0%	5%	0%	10%	0%	5%	0%	0%	21%	0%
Three	0%	0%	22%	0%	0%	0%	0%	0%	0%	0%	33%	5%	0%	0%	0%	0%	0%	5%	0%	5%	11%	5%	0%	5%	0%	0%	0%	0%	5%	0%
Turn	5%	0%	0%	0%	15%	0%	0%	0%	0%	20%	5%	0%	5%	0%	0%	0%	0%	0%	0%	0%	5%	0%	0%	0%	0%	10%	0%	35%	0%	0%
Two	0%	0%	0%	20%	0%	0%	0%	0%	0%	6%	33%	0%	0%	0%	0%	0%	6%	0%	0%	0%	0%	6%	0%	0%	13%	0%	6%	0%	6%	0%
Up	0%	0%	0%	0%	10%	0%	0%	0%	0%	5%	5%	5%	5%	0%	10%	0%	0%	5%	0%	0%	0%	5%	0%	0%	0%	0%	0%	50%	0%	0%
Yes	0%	0%	15%	0%	0%	0%	0%	0%	0%	0%	0%	25%	0%	0%	0%	0%	0%	5%	15%	0%	0%	10%	0%	0%	0%	0%	0%	0%	30%	0%
Zero	6%	0%	0%	0%	0%	0%	0%	0%	0%	0%	18%	12%	0%	12%	0%	0%	0%	6%	12%	0%	6%	0%	0%	0%	0%	6%	0%	0%	0%	18%

Tabla B.15. Ejemplo de Tabla B.de confusión del test de realizaciones ruidosas del experimento PLP-3b (PLP [10 coeficientes + 10 coeficientes delta <5 tramas> + 10 coeficientes delta-delta]).

	Double	Down	Eight	End	Five	Four	Go	Hit	Jump	Last	Left	Next	Nine	No	Off	One	On	Right	Seven	Six	Split	Stand	Stop	Ten	Three	Turn	Two	Up	Yes	Zero
Double	14%	0%	0%	0%	9%	9%	0%	0%	0%	0%	4%	4%	0%	14%	0%	4%	0%	9%	0%	0%	0%	4%	0%	0%	0%	4%	0%	4%	14%	0%
Down	4%	4%	0%	0%	13%	0%	4%	0%	0%	4%	4%	4%	4%	18%	0%	0%	0%	18%	0%	0%	0%	9%	0%	0%	0%	0%	0%	9%	0%	0%
Eight	0%	0%	47%	11%	0%	0%	0%	0%	0%	0%	0%	17%	0%	0%	0%	0%	0%	5%	5%	0%	0%	5%	0%	5%	0%	0%	0%	0%	0%	0%
End	0%	0%	0%	31%	18%	0%	0%	0%	0%	0%	6%	0%	0%	0%	0%	0%	0%	0%	6%	12%	0%	0%	0%	6%	0%	6%	0%	0%	12%	0%
Five	0%	0%	0%	0%	43%	0%	0%	0%	0%	4%	4%	4%	0%	0%	0%	4%	0%	21%	13%	0%	0%	0%	0%	0%	0%	0%	0%	4%	0%	0%
Four	0%	0%	12%	18%	6%	12%	0%	0%	0%	0%	6%	0%	0%	6%	12%	12%	0%	0%	0%	0%	0%	0%	0%	6%	0%	6%	0%	0%	0%	0%
Go	5%	0%	0%	11%	5%	0%	11%	0%	0%	0%	11%	5%	0%	22%	0%	5%	16%	5%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Hit	0%	0%	16%	0%	0%	0%	0%	0%	0%	0%	25%	25%	0%	0%	0%	0%	0%	0%	0%	8%	0%	8%	0%	0%	0%	0%	0%	0%	16%	0%
Jump	10%	0%	0%	0%	5%	0%	5%	0%	10%	10%	0%	5%	10%	10%	0%	0%	0%	0%	10%	0%	0%	0%	0%	0%	0%	5%	0%	15%	0%	0%
Last	5%	0%	0%	5%	15%	0%	0%	0%	0%	25%	0%	15%	10%	0%	0%	5%	0%	10%	0%	0%	0%	5%	0%	0%	0%	0%	0%	5%	0%	0%
Left	0%	0%	4%	0%	0%	0%	0%	0%	0%	4%	38%	23%	0%	0%	0%	0%	0%	0%	0%	4%	0%	9%	0%	4%	0%	0%	0%	0%	9%	0%
Next	0%	0%	0%	12%	0%	0%	0%	6%	0%	0%	18%	25%	0%	0%	0%	0%	0%	6%	0%	0%	0%	25%	0%	0%	0%	0%	0%	6%	0%	0%
Nine	0%	0%	0%	5%	11%	0%	0%	0%	0%	5%	11%	0%	22%	5%	0%	0%	0%	22%	0%	0%	0%	0%	0%	0%	0%	5%	0%	0%	11%	0%
No	22%	0%	0%	0%	5%	0%	5%	0%	0%	5%	0%	5%	0%	33%	0%	5%	0%	5%	0%	0%	0%	0%	0%	0%	0%	0%	0%	5%	0%	5%
Off	0%	0%	0%	0%	0%	7%	0%	0%	0%	0%	14%	7%	0%	0%	57%	7%	7%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
One	0%	0%	0%	0%	10%	0%	0%	0%	0%	5%	0%	0%	5%	0%	10%	5%	0%	47%	0%	0%	0%	0%	0%	5%	0%	0%	0%	10%	0%	0%
On	0%	0%	5%	17%	23%	0%	0%	0%	0%	0%	0%	0%	5%	0%	5%	29%	0%	11%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Right	0%	0%	0%	5%	15%	0%	0%	0%	0%	0%	5%	5%	0%	0%	0%	0%	0%	70%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Seven	0%	0%	0%	12%	0%	0%	0%	0%	0%	12%	18%	12%	0%	0%	0%	0%	0%	0%	18%	0%	0%	0%	0%	12%	0%	6%	0%	0%	6%	0%
Six	0%	0%	25%	0%	0%	0%	0%	0%	0%	0%	12%	25%	0%	0%	0%	0%	0%	0%	0%	25%	0%	0%	0%	12%	0%	0%	0%	0%	0%	0%
Split	0%	0%	13%	0%	6%	0%	0%	0%	0%	0%	6%	0%	0%	0%	0%	0%	0%	0%	6%	6%	6%	13%	0%	6%	13%	0%	0%	0%	20%	0%
Stand	0%	0%	0%	15%	36%	0%	0%	0%	0%	5%	0%	0%	5%	5%	0%	0%	0%	15%	5%	0%	0%	0%	0%	0%	0%	5%	0%	5%	0%	0%
Stop	6%	6%	0%	0%	12%	0%	0%	0%	0%	0%	18%	0%	0%	0%	12%	6%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	37%	0%	0%
Ten	0%	0%	5%	5%	5%	0%	0%	5%	0%	0%	21%	15%	0%	0%	0%	0%	0%	0%	0%	0%	0%	5%	0%	10%	0%	5%	0%	0%	21%	0%
Three	0%	0%	22%	0%	0%	0%	0%	0%	0%	0%	33%	5%	0%	0%	0%	0%	0%	5%	0%	5%	11%	5%	0%	5%	0%	0%	0%	0%	5%	0%
Turn	5%	0%	0%	0%	15%	0%	0%	0%	0%	20%	5%	0%	5%	0%	0%	0%	0%	0%	0%	0%	5%	0%	0%	0%	0%	10%	0%	35%	0%	0%
Two	0%	0%	0%	20%	0%	0%	0%	0%	0%	6%	33%	0%	0%	0%	0%	0%	6%	0%	0%	0%	0%	6%	0%	0%	13%	0%	6%	0%	6%	0%
Up	0%	0%	0%	0%	10%	0%	0%	0%	0%	5%	5%	5%	5%	0%	10%	0%	0%	5%	0%	0%	0%	5%	0%	0%	0%	0%	0%	50%	0%	0%
Yes	0%	0%	15%	0%	0%	0%	0%	0%	0%	0%	0%	25%	0%	0%	0%	0%	0%	5%	15%	0%	0%	10%	0%	0%	0%	0%	0%	0%	30%	0%
Zero	6%	0%	0%	0%	0%	0%	0%	0%	0%	0%	18%	12%	0%	12%	0%	0%	0%	6%	12%	0%	6%	0%	0%	0%	0%	6%	0%	0%	0%	18%

Tabla B.15. Ejemplo de Tabla B.de confusión del test de realizaciones ruidosas del experimento PLP-3b (PLP [10 coeficientes + 10 coeficientes delta <5 tramas> + 10 coeficientes delta-delta]).

APÉNDICE C. EXPERIMENTOS ACERCA DE LA INCORPORACIÓN DE CONOCIMIENTO ACÚSTICO

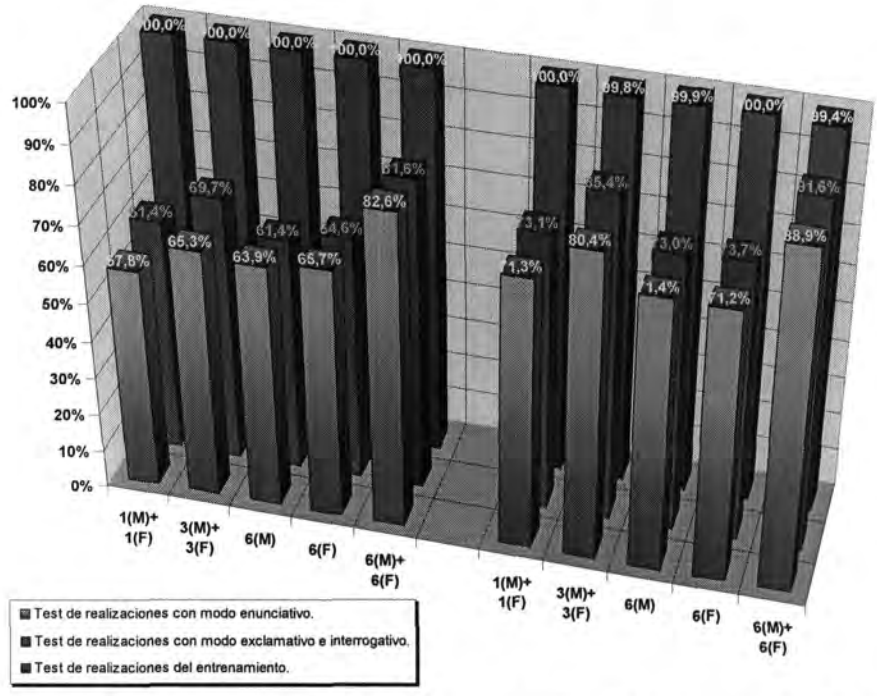

Figura C.1. Tasas de reconocimiento para el experimento FFT-8a (FFT-Cepstrum [10 coeficientes + 10 coeficientes delta <5 tramas> + 10 coeficientes delta-delta + energía trama + energía trama diferencia <2 tramas>]). La parte de la izquierda corresponde a los entrenamientos efectuados solamente con realizaciones con modo enunciativo.

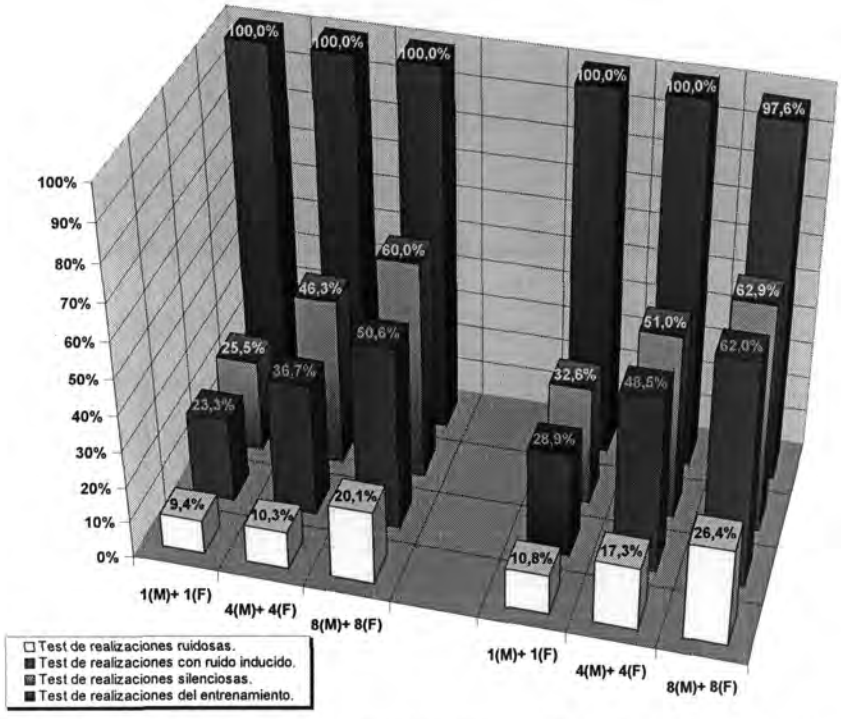

Figura C.2. Tasas de reconocimiento para el experimento FFT-8b (FFT-Cepstrum [10 coeficientes + 10 coeficientes delta <5 tramas> + 10 coeficientes delta-delta + energía trama + energía trama diferencia <2 tramas>]). La parte de la izquierda corresponde a los entrenamientos efectuados solamente con realizaciones silenciosas.

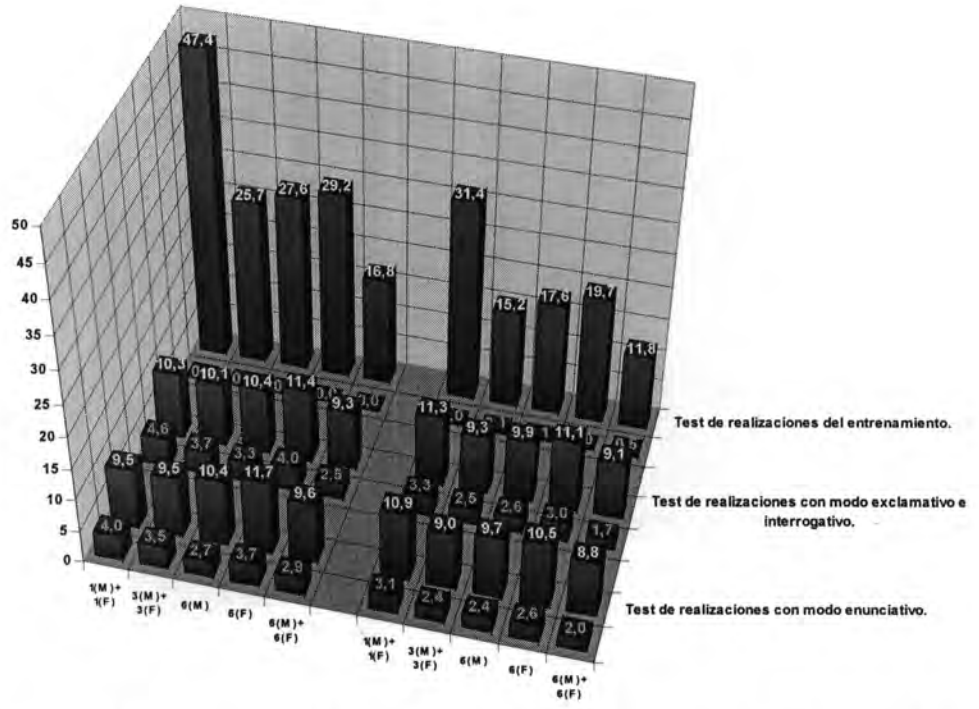

Figura C.3. Tasas de fiabilidad para el experimento FFT-8a (FFT-Cepstrum [10 coeficientes + 10 coeficientes delta <5 tramas> + 10 coeficientes delta-delta + energía trama + energía trama diferencia <2 tramas>]). En verde aparecen las tasas para los casos reconocidos correctamente. En rojo los casos clasificados erróneamente.

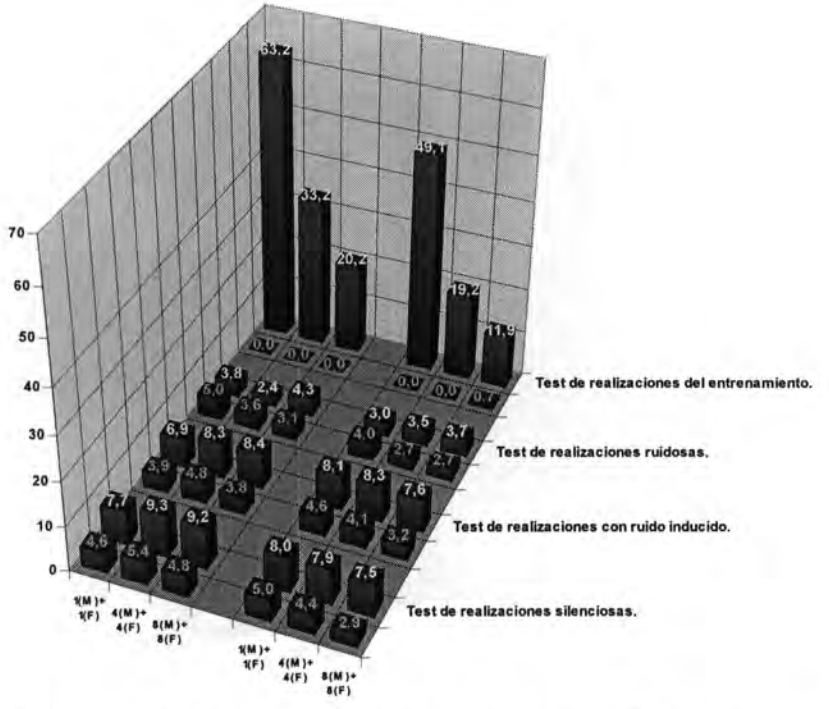

Figura C.4. Tasas de fiabilidad para el experimento FFT-8b (FFT-Cepstrum [10 coeficientes + 10 coeficientes delta <5 tramas> + 10 coeficientes delta-delta + energía trama + energía trama diferencia <2 tramas>]). En verde aparecen las tasas para los casos reconocidos correctamente. En rojo los casos clasificados erróneamente.

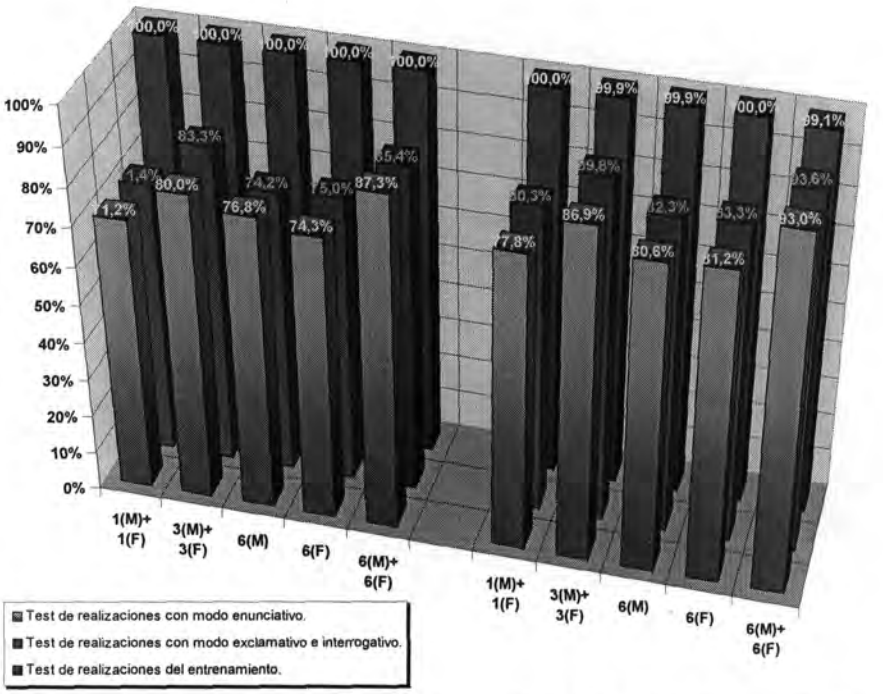

Figura C.5. Tasas de reconocimiento para el experimento FFT-$8_{(1)}$a (FFT-Cepstrum [6 coeficientes + 6 coeficientes delta <5 tramas>]). La parte de la izquierda corresponde a los entrenamientos efectuados solamente con realizaciones con modo enunciativo.

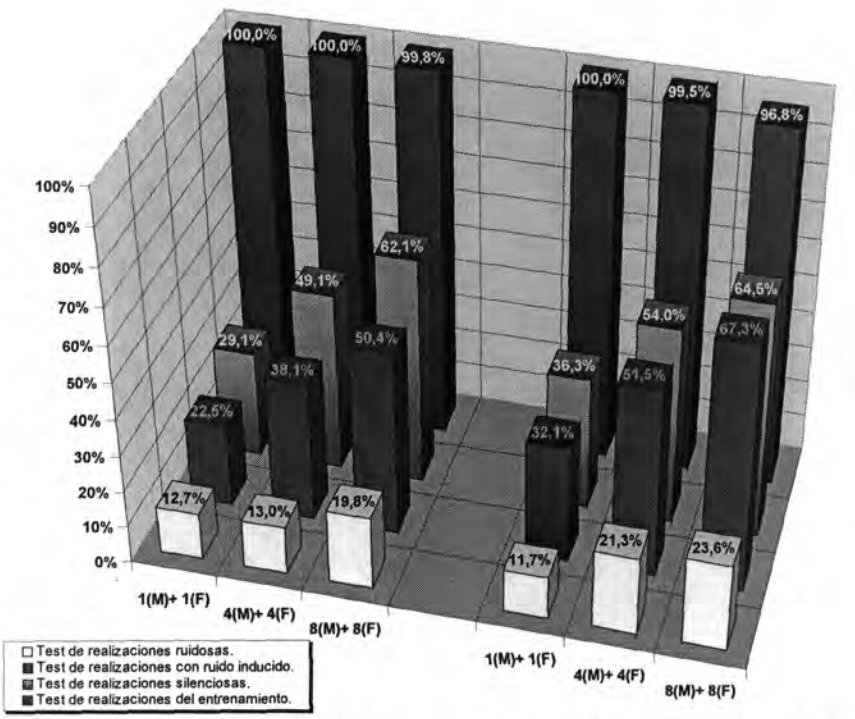

Figura C.6. Tasas de reconocimiento para el experimento FFT-$8_{(1)}$b (FFT-Cepstrum [6 coeficientes + 6 coeficientes delta <5 tramas>]). La parte de la izquierda corresponde a los entrenamientos efectuados solamente con realizaciones silenciosas.

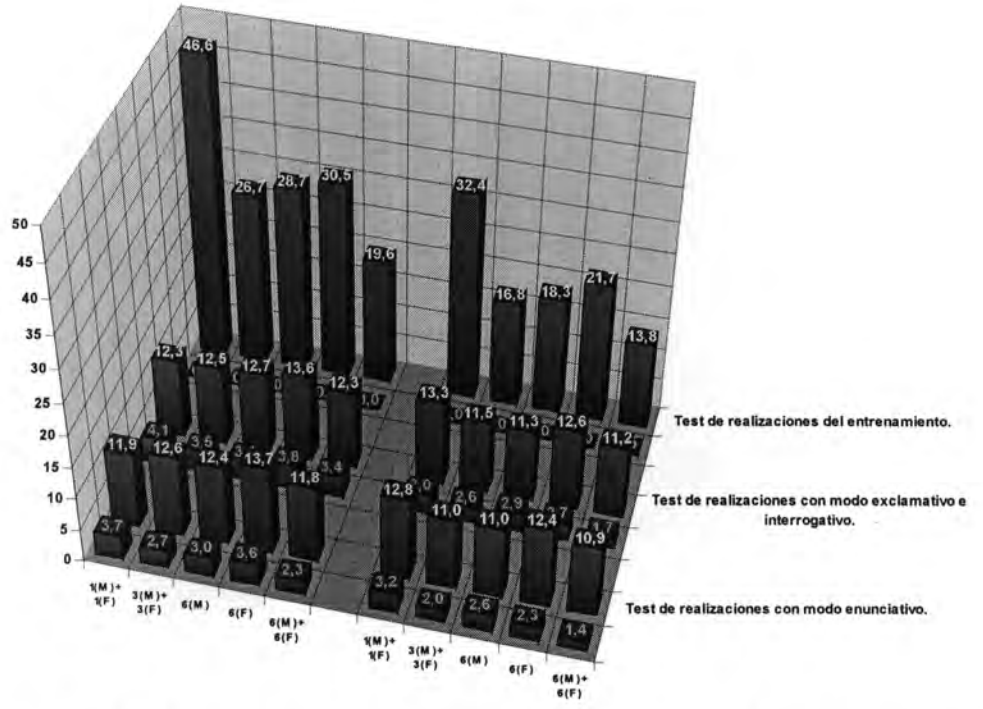

Figura C.7. Tasas de fiabilidad para el experimento FFT-$8_{(1)}$a (FFT-Cepstrum [6 coeficientes + 6 coeficientes delta <5 tramas>]). En verde aparecen las tasas para los casos reconocidos correctamente. En rojo los casos clasificados erróneamente.

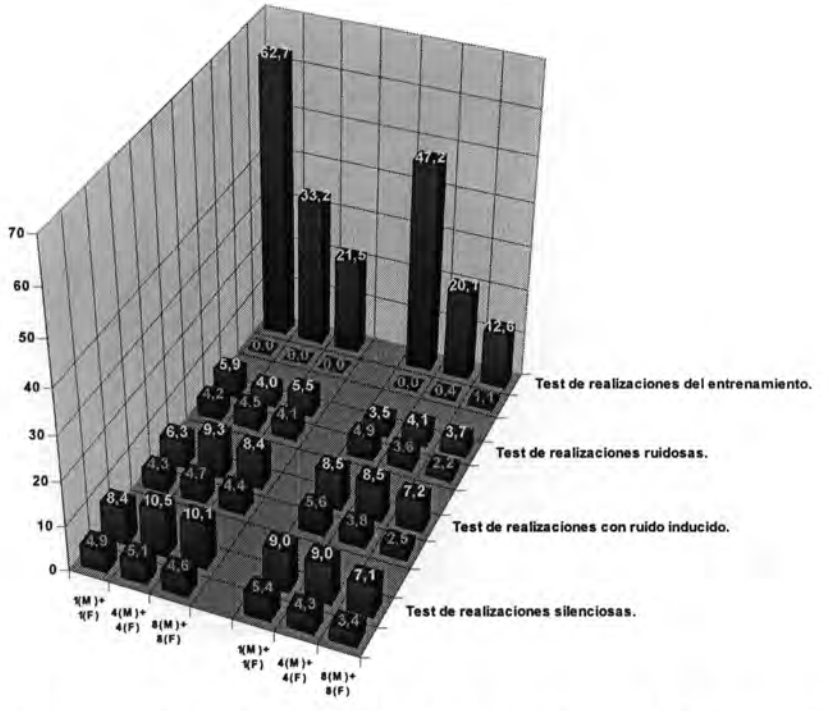

Figura C.8. Tasas de fiabilidad para el experimento FFT-$8_{(1)}$b (FFT-Cepstrum [6 coeficientes + 6 coeficientes delta <5 tramas>]). En verde aparecen las tasas para los casos reconocidos correctamente. En rojo los casos clasificados erróneamente.

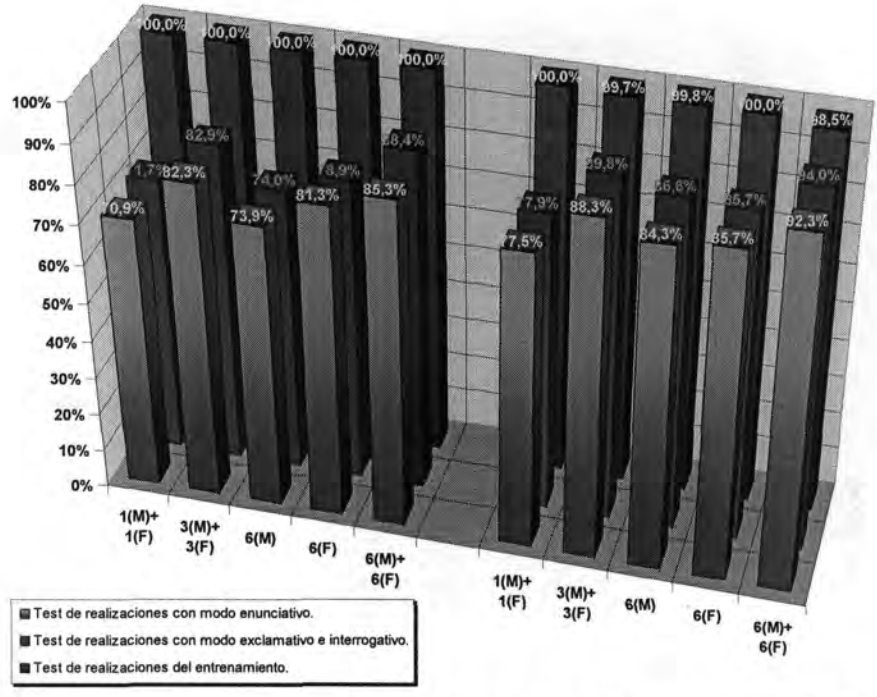

Figura C.9. Tasas de reconocimiento para el experimento FFT-$8_{(2)}$a (FFT-Cepstrum [4 coeficientes + 4 coeficientes delta <5 tramas>]). La parte de la izquierda corresponde a los entrenamientos efectuados solamente con realizaciones con modo enunciativo.

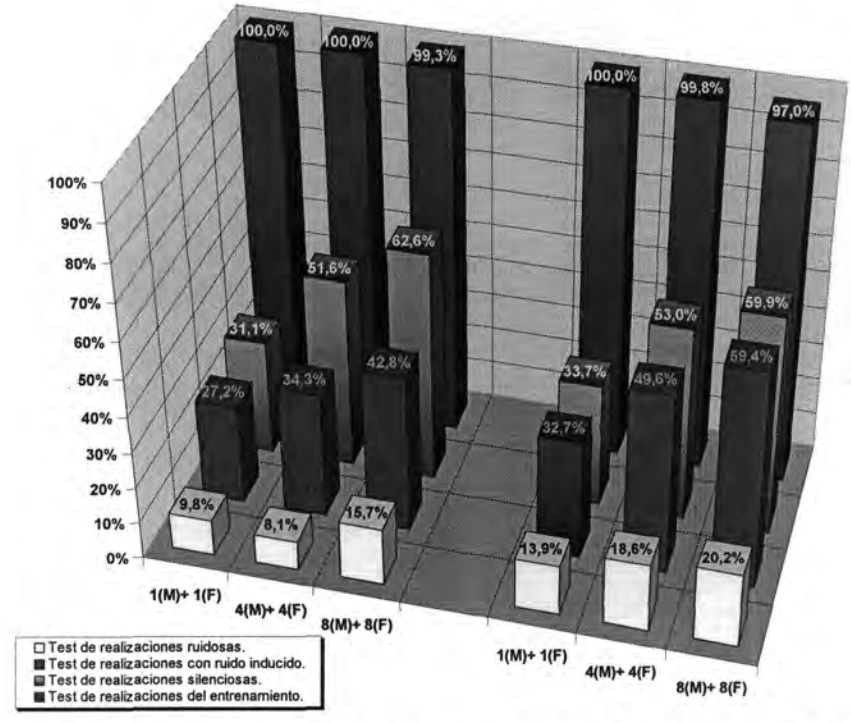

Figura C.10. Tasas de reconocimiento para el experimento FFT-$8_{(2)}$b (FFT-Cepstrum [4 coeficientes + 4 coeficientes delta <5 tramas>]). La parte de la izquierda corresponde a los entrenamientos efectuados solamente con realizaciones silenciosas.

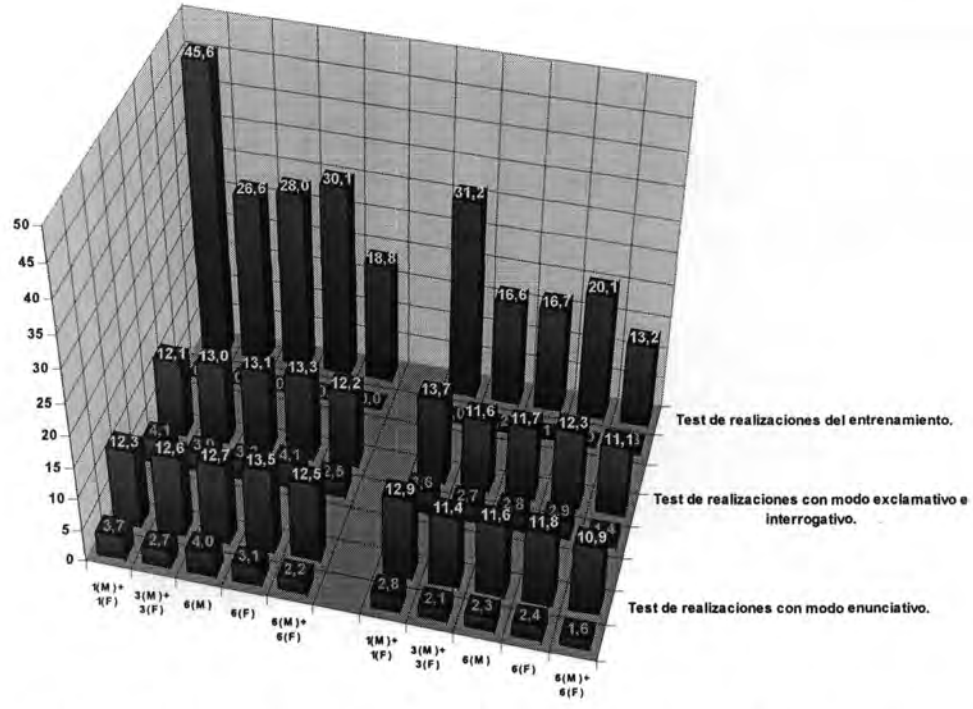

Figura C.11. Tasas de fiabilidad para el experimento FFT-$8_{(2)}$a (FFT-Cepstrum [4 coeficientes + 4 coeficientes delta <5 tramas>]). En verde aparecen las tasas para los casos reconocidos correctamente. En rojo los casos clasificados erróneamente.

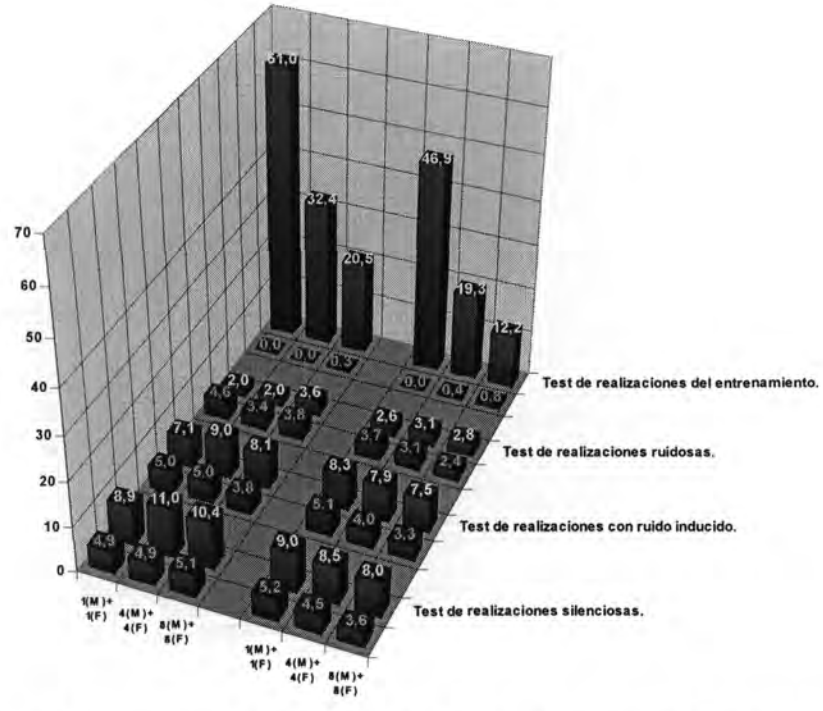

Figura C.12. Tasas de fiabilidad para el experimento FFT-$8_{(2)}$b (FFT-Cepstrum [4 coeficientes + 4 coeficientes delta <5 tramas>]). En verde aparecen las tasas para los casos reconocidos correctamente. En rojo los casos clasificados erróneamente.

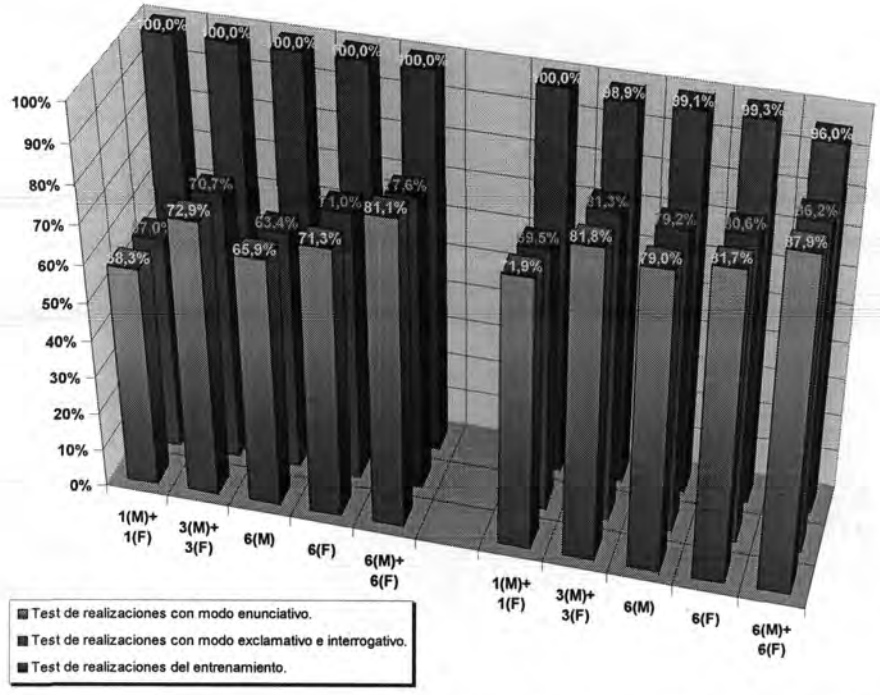

Figura C.13. Tasas de reconocimiento para el experimento FFT-$8_{(3)}$a (FFT-Cepstrum [2 coeficientes + 2 coeficientes delta <5 tramas>]). La parte de la izquierda corresponde a los entrenamientos efectuados solamente con realizaciones con modo enunciativo.

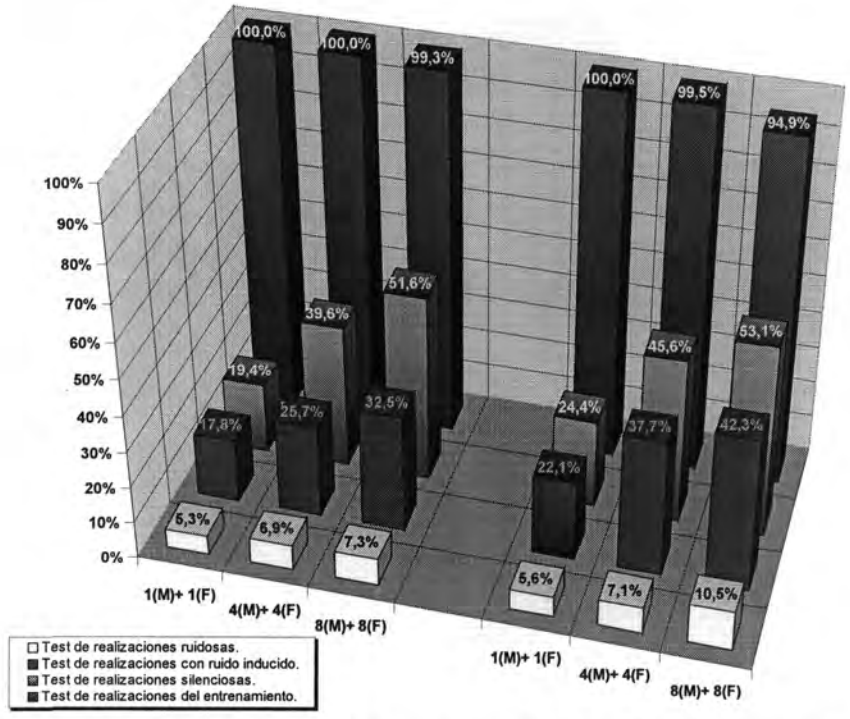

Figura C.14. Tasas de reconocimiento para el experimento FFT-$8_{(3)}$b (FFT-Cepstrum [2 coeficientes + 2 coeficientes delta <5 tramas>]). La parte de la izquierda corresponde a los entrenamientos efectuados solamente con realizaciones silenciosas.

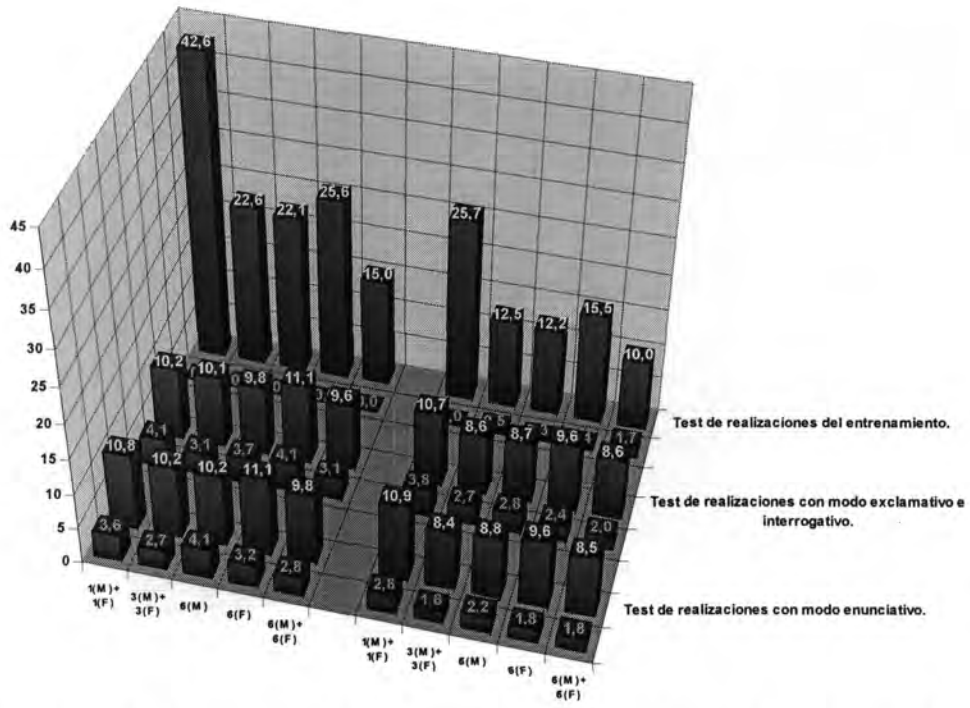

Figura C.15. Tasas de fiabilidad para el experimento FFT-$8_{(3)}$a (FFT-Cepstrum [2 coeficientes + 2 coeficientes delta <5 tramas>]). En verde aparecen las tasas para los casos reconocidos correctamente. En rojo los casos clasificados erróneamente.

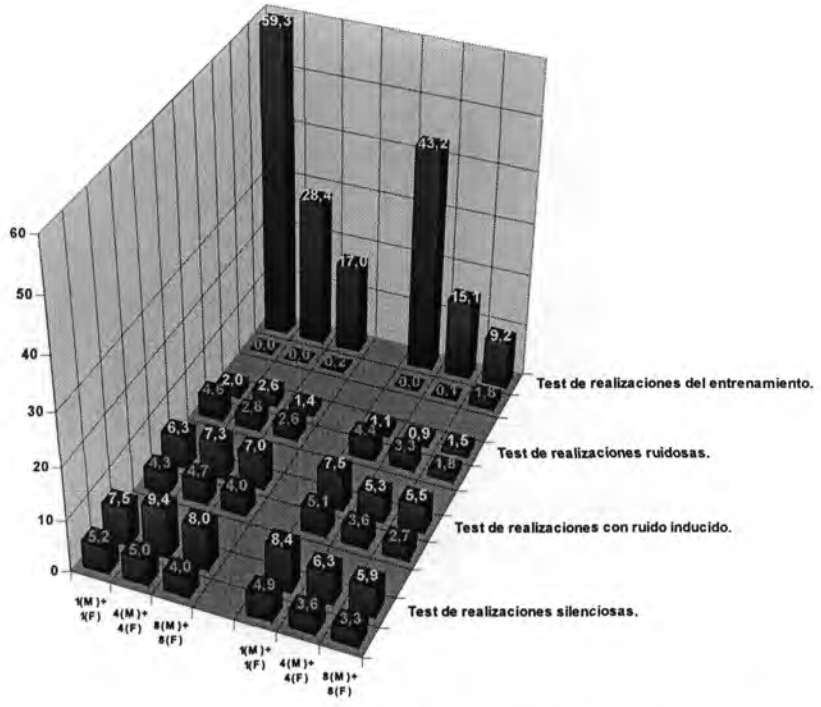

Figura C.16. Tasas de fiabilidad para el experimento FFT-$8_{(3)}$b (FFT-Cepstrum [2 coeficientes + 2 coeficientes delta <5 tramas>]). En verde aparecen las tasas para los casos reconocidos correctamente. En rojo los casos clasificados erróneamente.

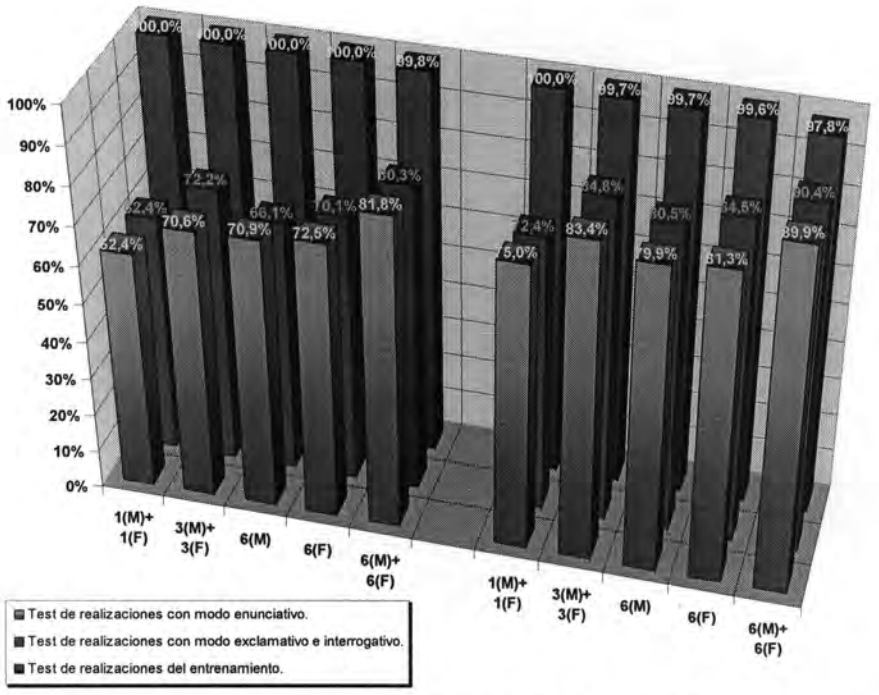

Figura C.17. Tasas de reconocimiento para el experimento FFT-$8_{(4)}$a (FFT-Cepstrum [10 coeficientes + 10 coeficientes delta <5 tramas>]. Cuantificación en 2 niveles: nivel 1° <16 índices> [6 coeficientes + 6 coeficientes delta], nivel 2° <240 índices> [4 coeficientes + 4 coeficientes delta]). La parte de la izquierda corresponde a los entrenamientos efectuados solamente con realizaciones con modo enunciativo.

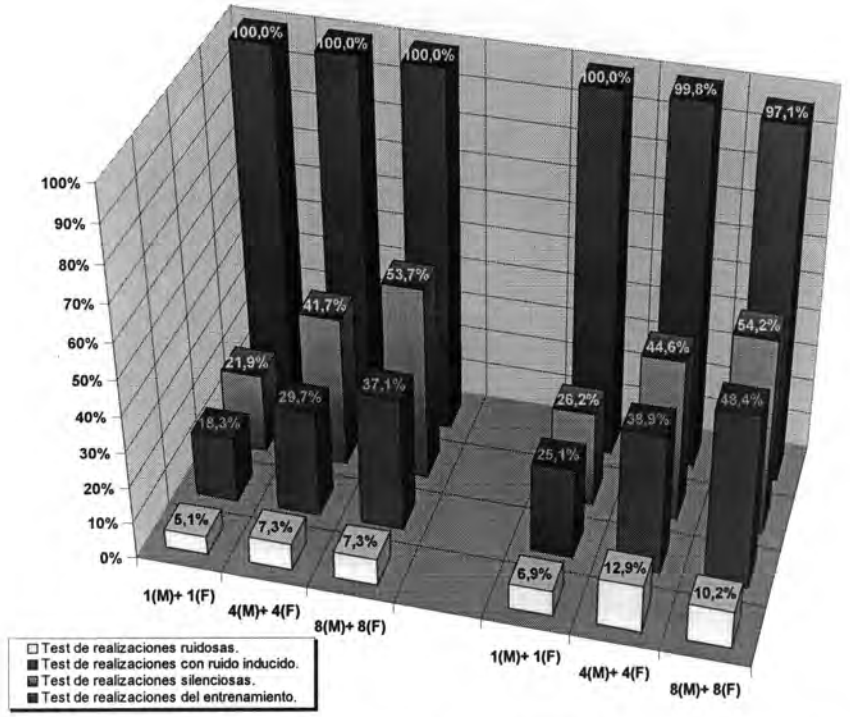

Figura C.18. Tasas de reconocimiento para el experimento FFT-$8_{(4)}$b (FFT-Cepstrum [10 coeficientes + 10 coeficientes delta <5 tramas>]. Cuantificación en 2 niveles: nivel 1° <16 índices> [6 coeficientes + 6 coeficientes delta], nivel 2° <240 índices> [4 coeficientes + 4 coeficientes delta]). La parte de la izquierda corresponde a los entrenamientos efectuados solamente con realizaciones silenciosas.

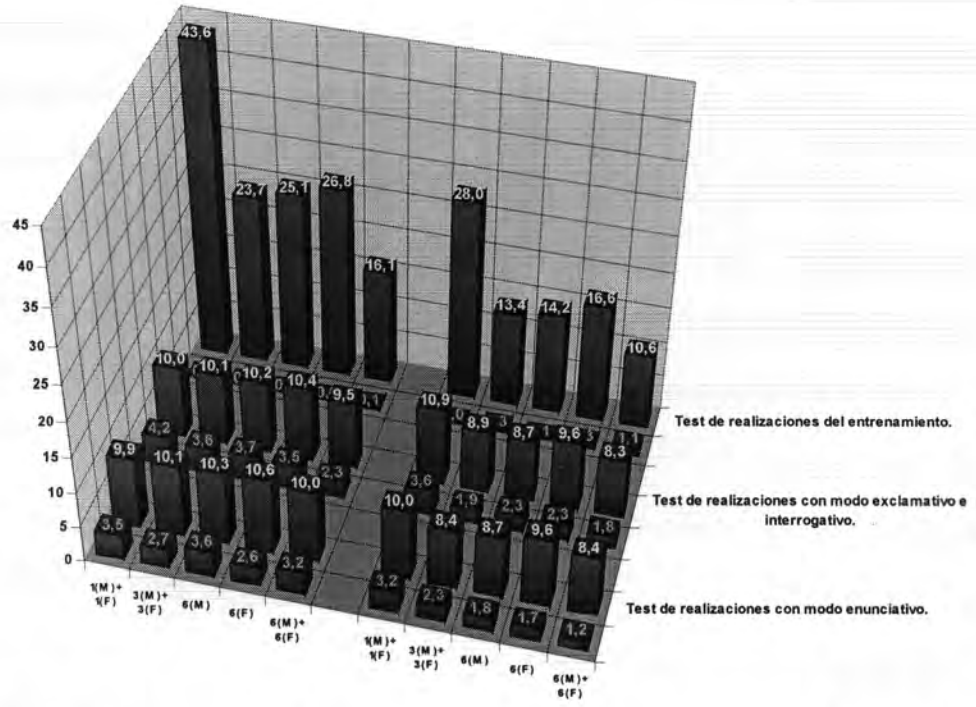

Figura C.19. Tasas de fiabilidad para el experimento FFT-$8_{(4)}$a (FFT-Cepstrum [10 coeficientes + 10 coeficientes delta <5 tramas>]. Cuantificación en 2 niveles: nivel 1º <16 índices> [6 coeficientes + 6 coeficientes delta], nivel 2º <240 índices> [4 coeficientes + 4 coeficientes delta]). En verde aparecen las tasas para los casos reconocidos correctamente. En rojo los casos clasificados erróneamente.

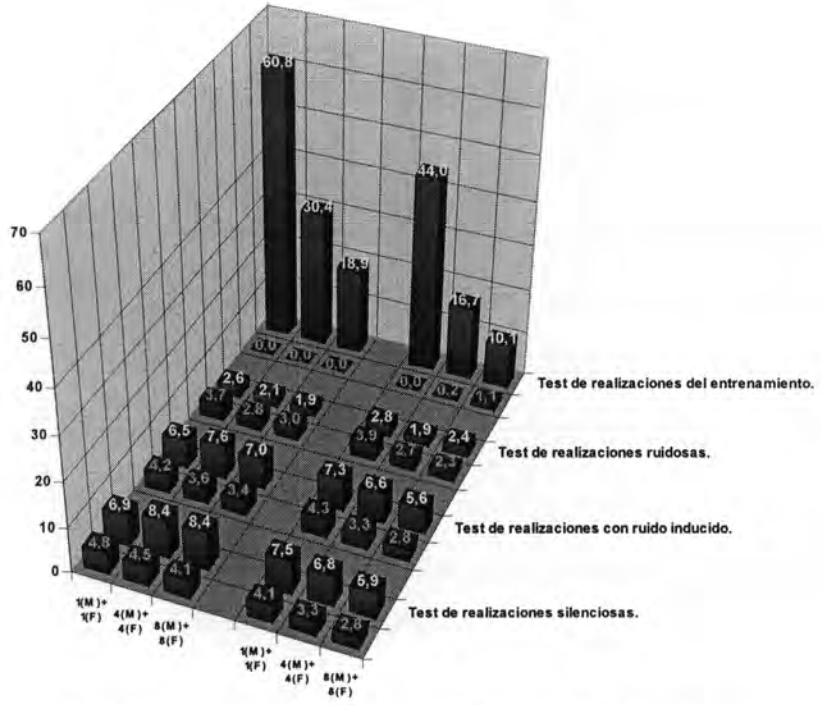

Figura C.20. Tasas de fiabilidad para el experimento FFT-$8_{(4)}$b (FFT-Cepstrum [10 coeficientes + 10 coeficientes delta <5 tramas>]. Cuantificación en 2 niveles: nivel 1º <16 índices> [6 coeficientes + 6 coeficientes delta], nivel 2º <240 índices> [4 coeficientes + 4 coeficientes delta]). En verde aparecen las tasas para los casos reconocidos correctamente. En rojo los casos clasificados erróneamente.

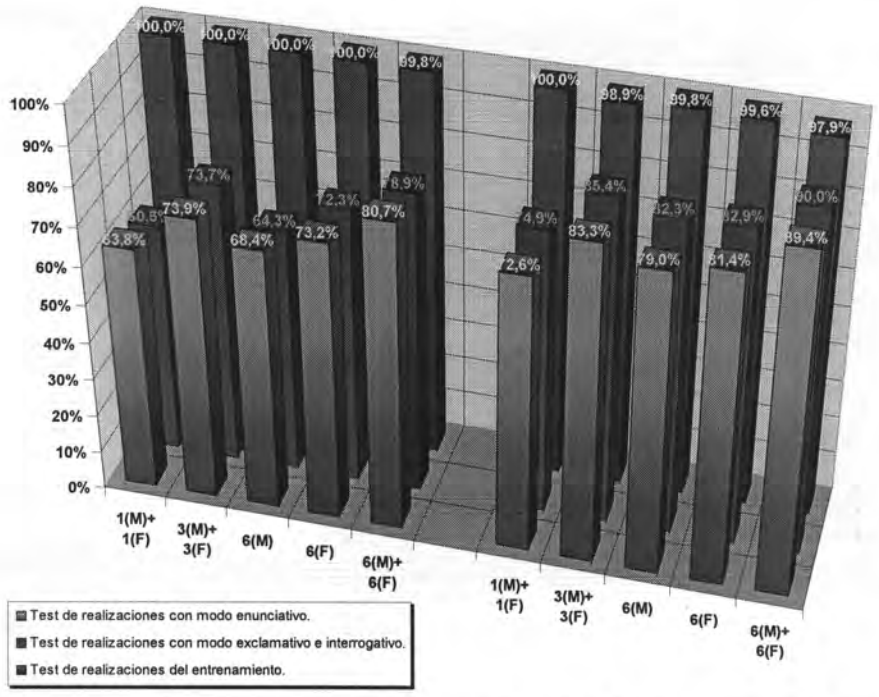

Figura C.21. Tasas de reconocimiento para el experimento FFT-$8_{(5)}$a (FFT-Cepstrum [10 coeficientes + 10 coeficientes delta <5 tramas>]. Cuantificación en 3 niveles: nivel 1° <16 índices> [4 coeficientes + 4 coeficientes delta], nivel 2° <48 índices> [2 coeficientes + 2 coeficientes delta], nivel 3° <192 índices> [4 coeficientes + 4 coeficientes delta]). La parte de la izquierda corresponde a los entrenamientos efectuados solamente con realizaciones con modo enunciativo.

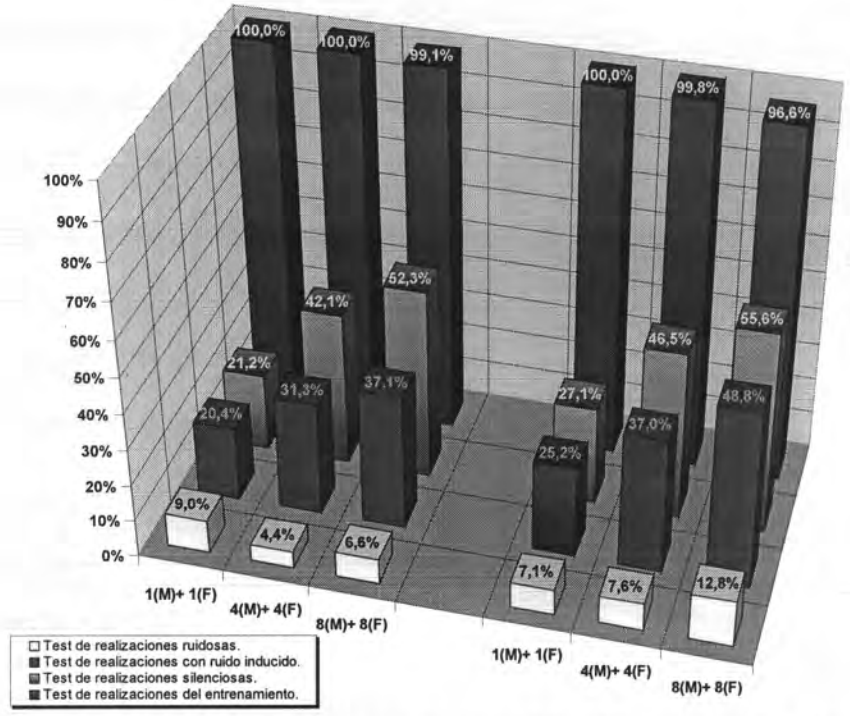

Figura C.22. Tasas de reconocimiento para el experimento FFT-$8_{(5)}$b (FFT-Cepstrum [10 coeficientes + 10 coeficientes delta <5 tramas>]. Cuantificación en 3 niveles: nivel 1° <16 índices> [4 coeficientes + 4 coeficientes delta], nivel 2° <48 índices> [2 coeficientes + 2 coeficientes delta], nivel 3° <192 índices> [4 coeficientes + 4 coeficientes delta]). La parte de la izquierda corresponde a los entrenamientos efectuados solamente con realizaciones silenciosas.

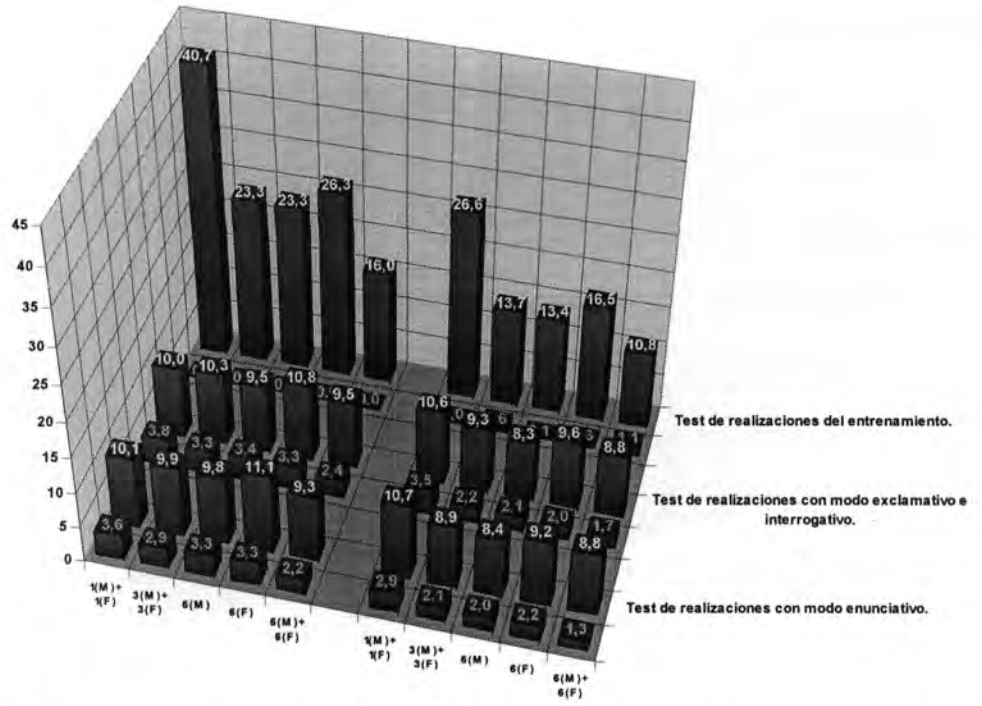

Figura C.23. Tasas de fiabilidad para el experimento FFT-$8_{(5)}$a (FFT-Cepstrum [10 coeficientes + 10 coeficientes delta <5 tramas>]. Cuantificación en 3 niveles: nivel 1º <16 índices> [4 coeficientes + 4 coeficientes delta], nivel 2º <48 índices> [2 coeficientes + 2 coeficientes delta], nivel 3º <192 índices> [4 coeficientes + 4 coeficientes delta]). En verde aparecen las tasas para los casos reconocidos correctamente. En rojo los casos clasificados erróneamente.

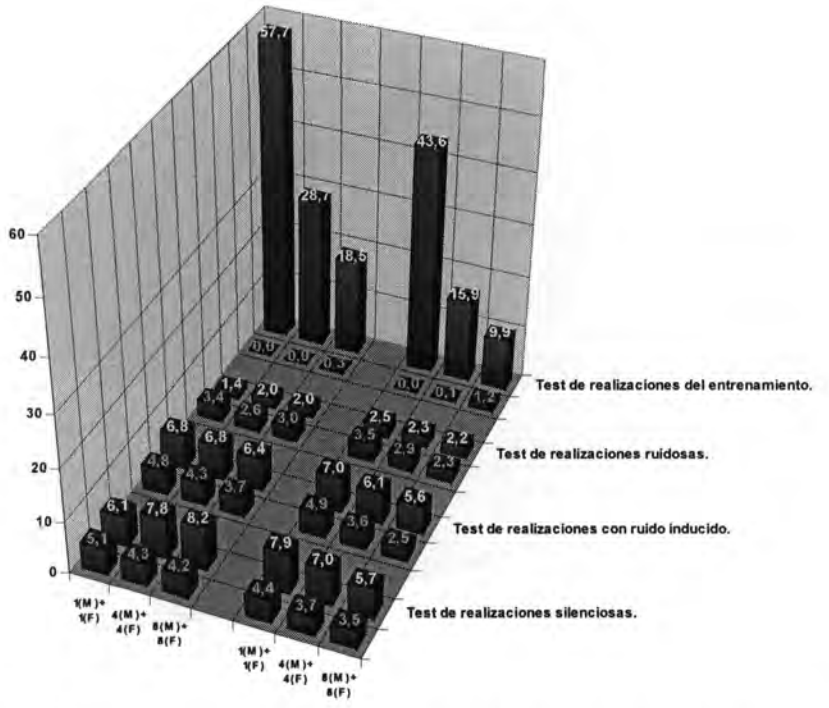

Figura C.24. Tasas de fiabilidad para el experimento FFT-$8_{(5)}$b (FFT-Cepstrum [10 coeficientes + 10 coeficientes delta <5 tramas>]. Cuantificación en 3 niveles: nivel 1º <16 índices> [4 coeficientes + 4 coeficientes delta], nivel 2º <48 índices> [2 coeficientes + 2 coeficientes delta], nivel 3º <192 índices> [4 coeficientes + 4 coeficientes delta]). En verde aparecen las tasas para los casos reconocidos correctamente. En rojo los casos clasificados erróneamente.

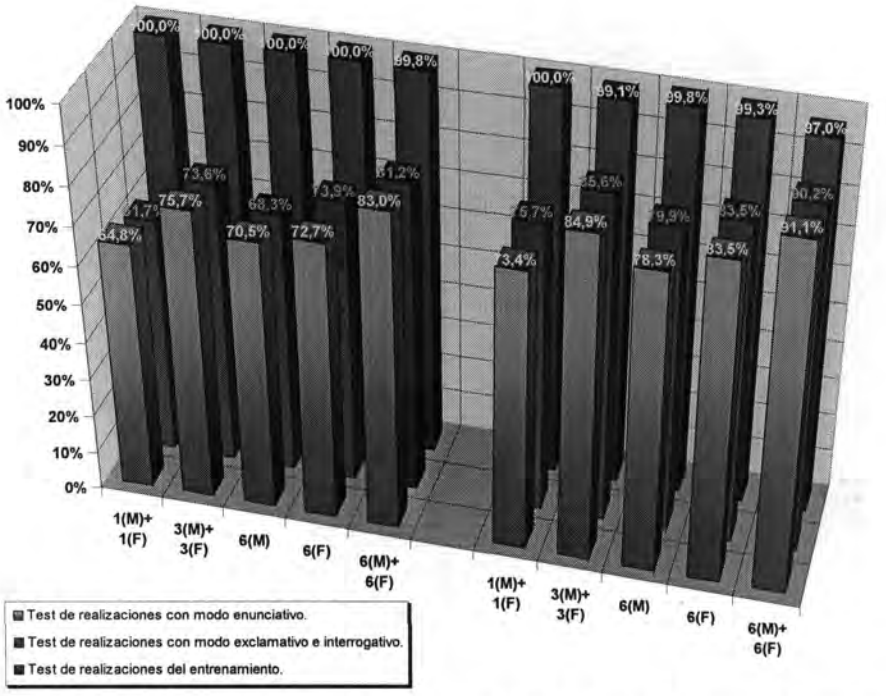

Figura C.25. Tasas de reconocimiento para el experimento FFT-$8_{(6)}$a (FFT-Cepstrum [10 coeficientes + 10 coeficientes delta <5 tramas>]. Cuantificación en 2 niveles: nivel 1º <16 índices> [6 coeficientes + 6 coeficientes delta], nivel 2º <240 índices> [4 coeficientes + 4 coeficientes delta]). La parte de la izquierda corresponde a los entrenamientos efectuados solamente con realizaciones con modo enunciativo.

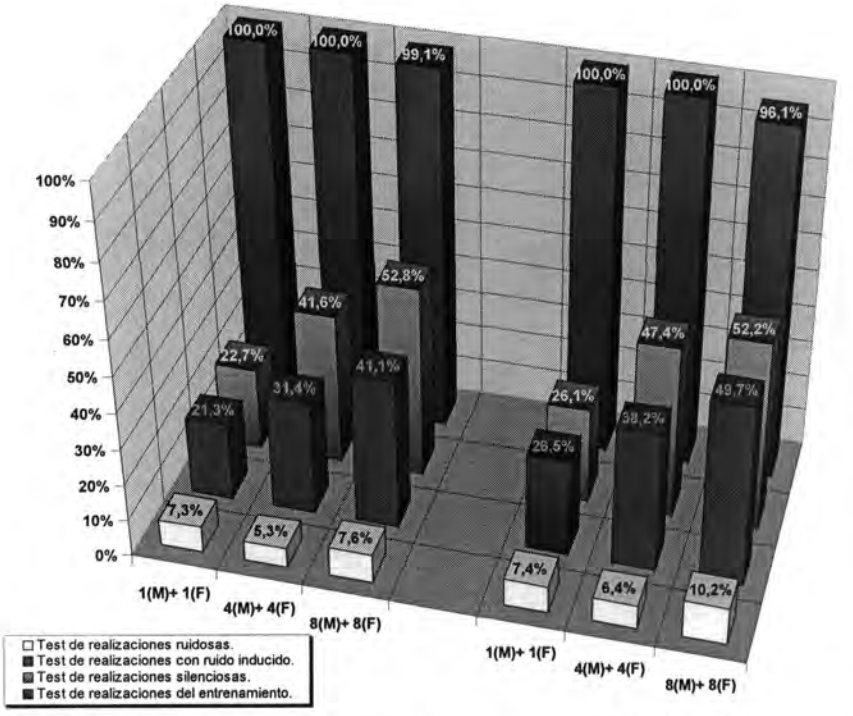

Figura C.26. Tasas de reconocimiento para el experimento FFT-$8_{(6)}$b (FFT-Cepstrum [10 coeficientes + 10 coeficientes delta <5 tramas>]. Cuantificación en 4 niveles: nivel 1º <16 índices> [2 coeficientes + 2 coeficientes delta], nivel 2º <48 índices> [2 coeficientes + 2 coeficientes delta], nivel 3º <64 índices> [2 coeficientes + 2 coeficientes delta], nivel 4º <128 índices> [4 coeficientes + 4 coeficientes delta]). La parte de la izquierda corresponde a los entrenamientos efectuados solamente con realizaciones silenciosas.

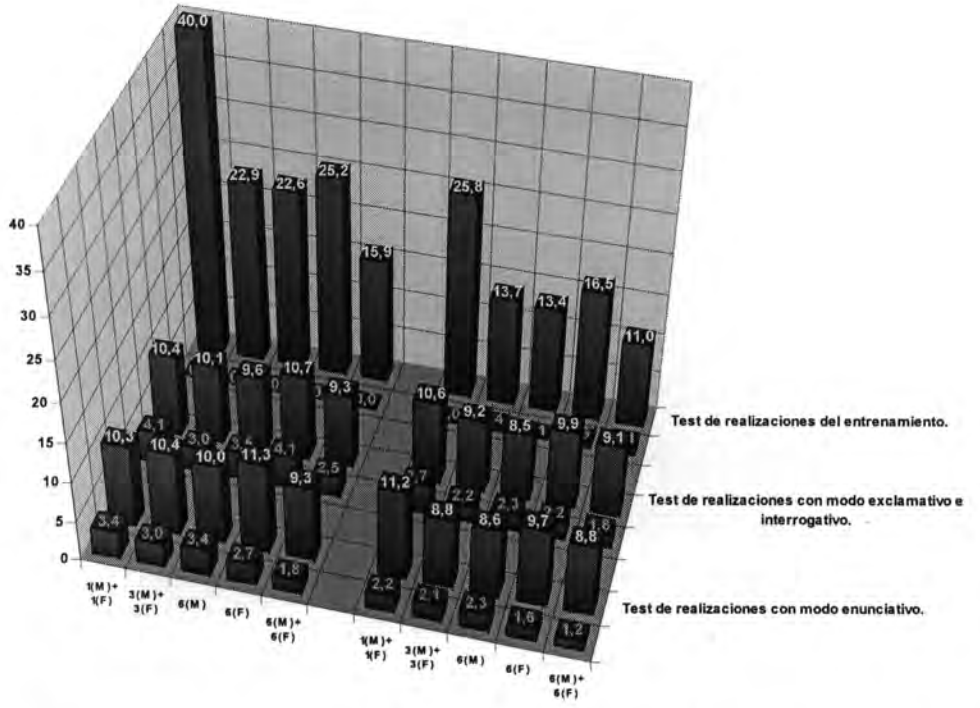

Figura C.27. Tasas de fiabilidad para el experimento FFT-$8_{(6)}$a (FFT-Cepstrum [10 coeficientes + 10 coeficientes delta <5 tramas>]. Cuantificación en 4 niveles: nivel 1º <16 índices> [2 coeficientes + 2 coeficientes delta], nivel 2º <48 índices> [2 coeficientes + 2 coeficientes delta], nivel 3º <64 índices> [2 coeficientes + 2 coeficientes delta], nivel 4º <128 índices> [4 coeficientes + 4 coeficientes delta]). En verde aparecen las tasas para los casos reconocidos correctamente. En rojo los casos clasificados erróneamente.

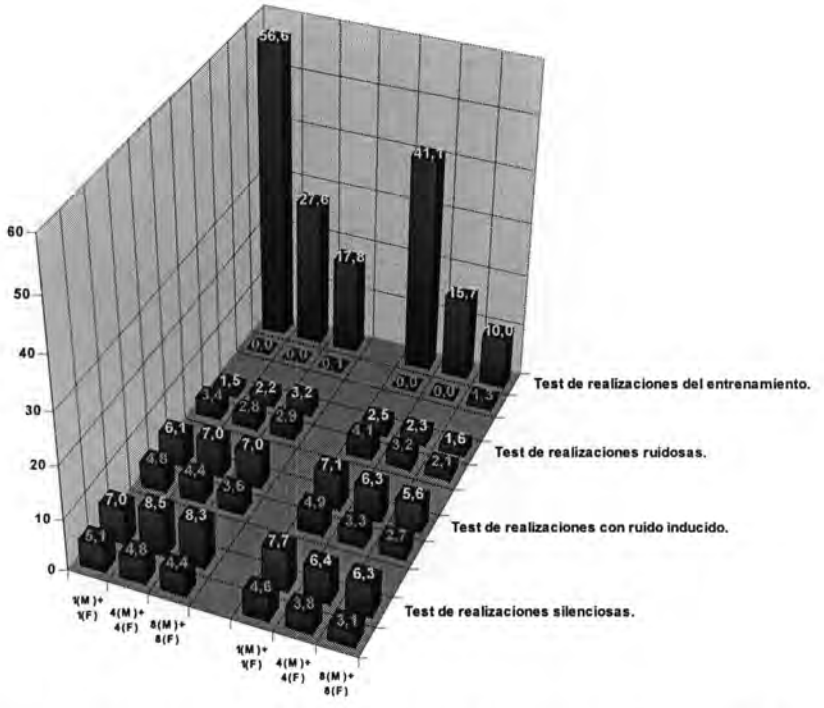

Figura C.28. Tasas de fiabilidad para el experimento FFT-$8_{(6)}$b (FFT-Cepstrum [10 coeficientes + 10 coeficientes delta <5 tramas>]. Cuantificación en 4 niveles: nivel 1º <16 índices> [2 coeficientes + 2 coeficientes delta], nivel 2º <48 índices> [2 coeficientes + 2 coeficientes delta], nivel 3º <64 índices> [2 coeficientes + 2 coeficientes delta], nivel 4º <128 índices> [4 coeficientes + 4 coeficientes delta]). En verde aparecen las tasas para los casos reconocidos correctamente. En rojo los casos clasificados erróneamente.

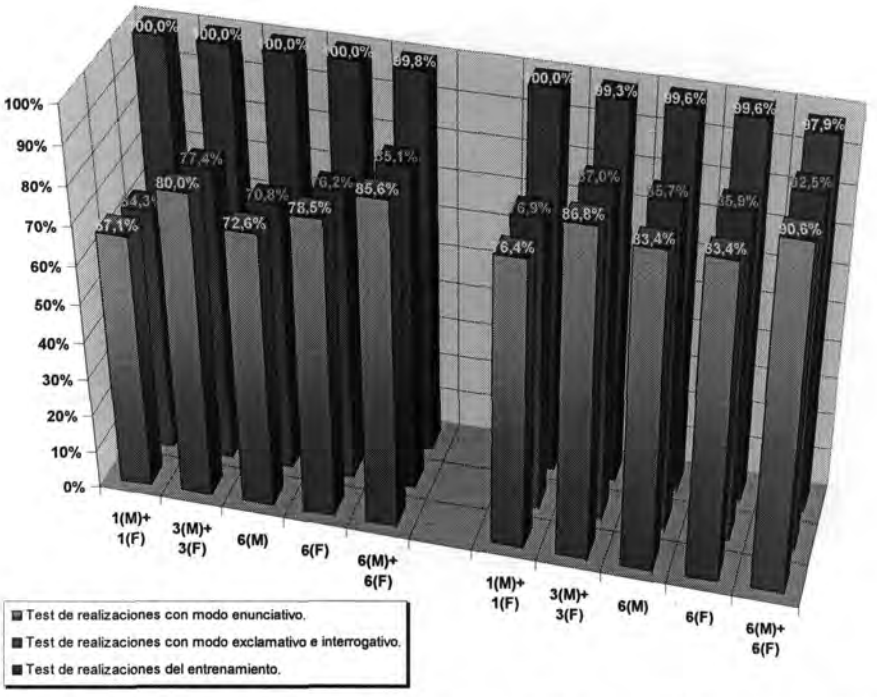

Figura C.29. Tasas de reconocimiento para el experimento FFT-$8_{(7)}$a (FFT-Cepstrum [4 coeficientes + 4 coeficientes delta <5 tramas>]. Cuantificación en 2 niveles: nivel 1º <64 índices> [2 coeficientes + 2 coeficientes delta], nivel 2º <192 índices> [2 coeficientes + 2 coeficientes delta]). La parte de la izquierda corresponde a los entrenamientos efectuados solamente con realizaciones con modo enunciativo.

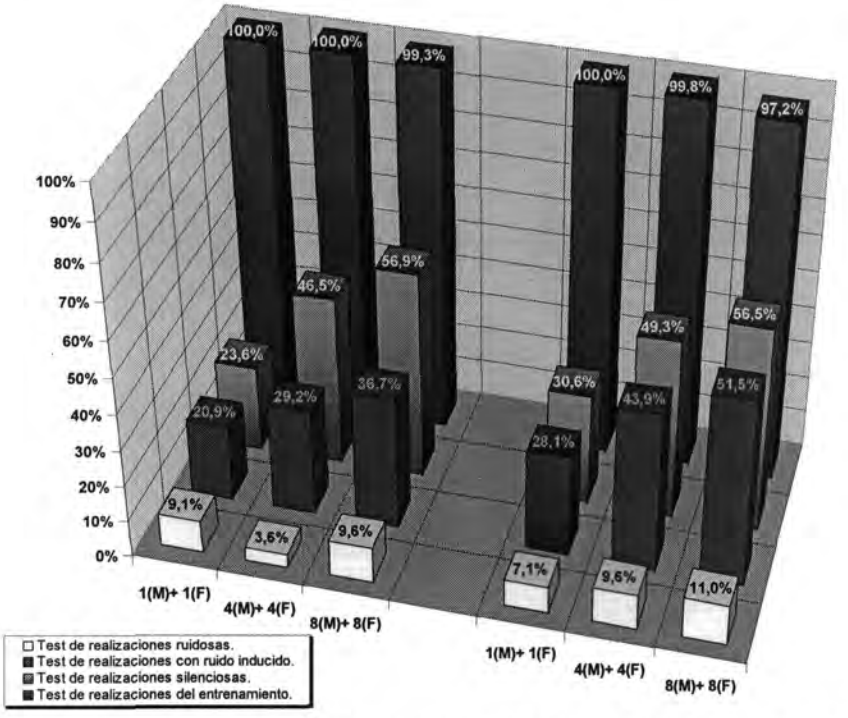

Figura C.30. Tasas de reconocimiento para el experimento FFT-$8_{(7)}$b (FFT-Cepstrum [4 coeficientes + 4 coeficientes delta <5 tramas>]. Cuantificación en 2 niveles: nivel 1º <64 índices> [2 coeficientes + 2 coeficientes delta], nivel 2º <192 índices> [2 coeficientes + 2 coeficientes delta]). La parte de la izquierda corresponde a los entrenamientos efectuados solamente con realizaciones silenciosas.

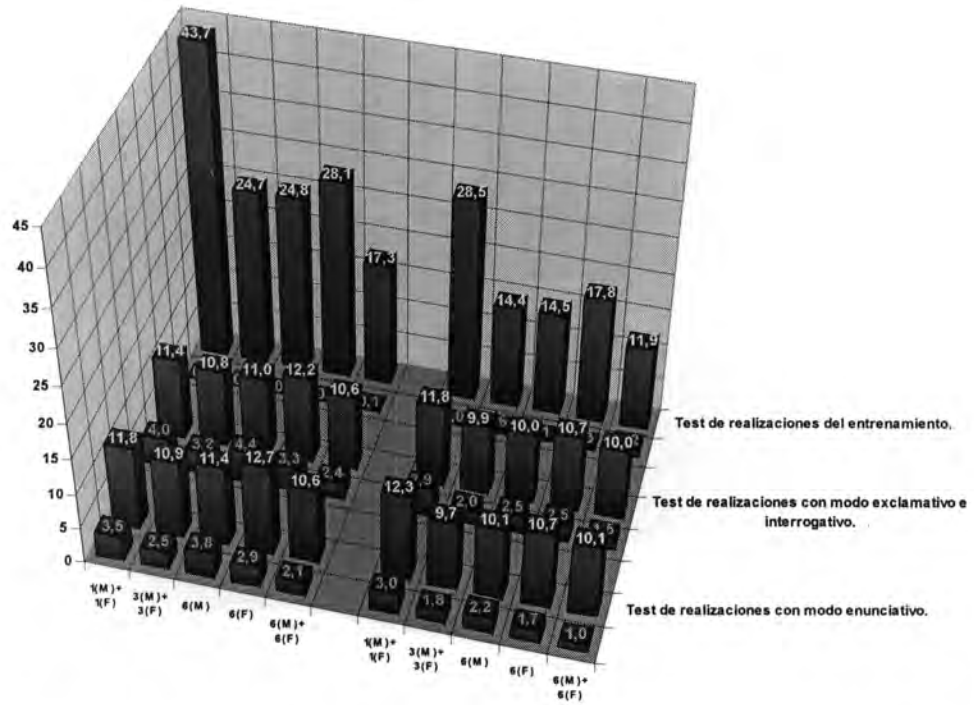

Figura C.31. Tasas de fiabilidad para el experimento FFT-8$_{(7)}$a (FFT-Cepstrum [4 coeficientes + 4 coeficientes delta <5 tramas>]. Cuantificación en 2 niveles: nivel 1º <64 índices> [2 coeficientes + 2 coeficientes delta], nivel 2º <192 índices> [2 coeficientes + 2 coeficientes delta]). En verde aparecen las tasas para los casos reconocidos correctamente. En rojo los casos clasificados erróneamente.

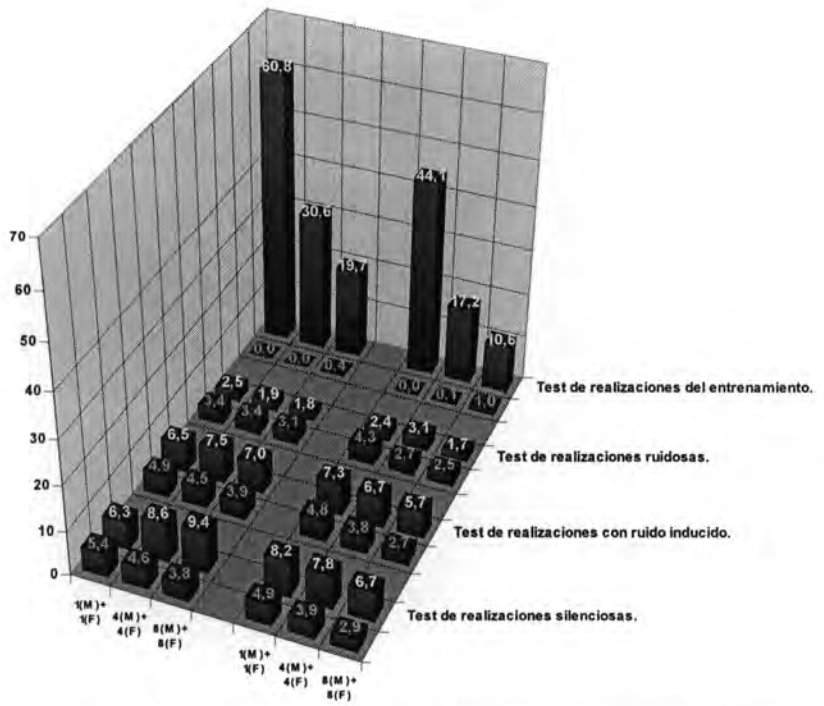

Figura C.32. Tasas de fiabilidad para el experimento FFT-8$_{(7)}$b (FFT-Cepstrum [4 coeficientes + 4 coeficientes delta <5 tramas>]. Cuantificación en 2 niveles: nivel 1º <64 índices> [2 coeficientes + 2 coeficientes delta], nivel 2º <192 índices> [2 coeficientes + 2 coeficientes delta]). En verde aparecen las tasas para los casos reconocidos correctamente. En rojo los casos clasificados erróneamente.

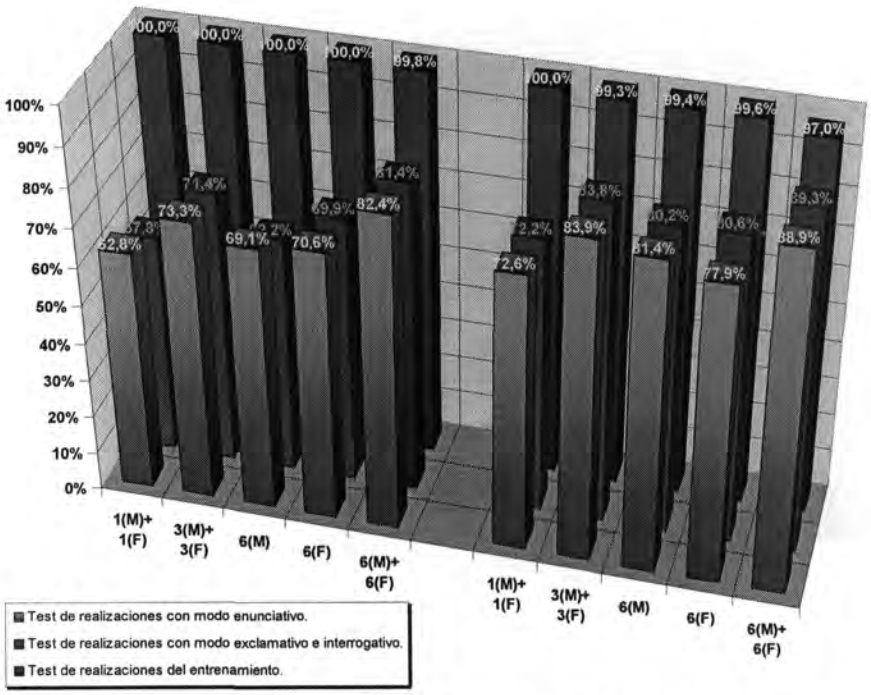

Figura C.33. Tasas de reconocimiento para el experimento FFT-$8_{(8)}$a (FFT-Cepstrum [4 coeficientes + 4 coeficientes delta <5 tramas>]. Cuantificación en 3 niveles: nivel 1° <16 índices> [1 coeficiente + 1 coeficiente delta], nivel 2° <48 índices> [1 coeficiente + 1 coeficiente delta], nivel 3° <192 índices> [2 coeficientes + 2 coeficientes delta]). La parte de la izquierda corresponde a los entrenamientos efectuados solamente con realizaciones con modo enunciativo.

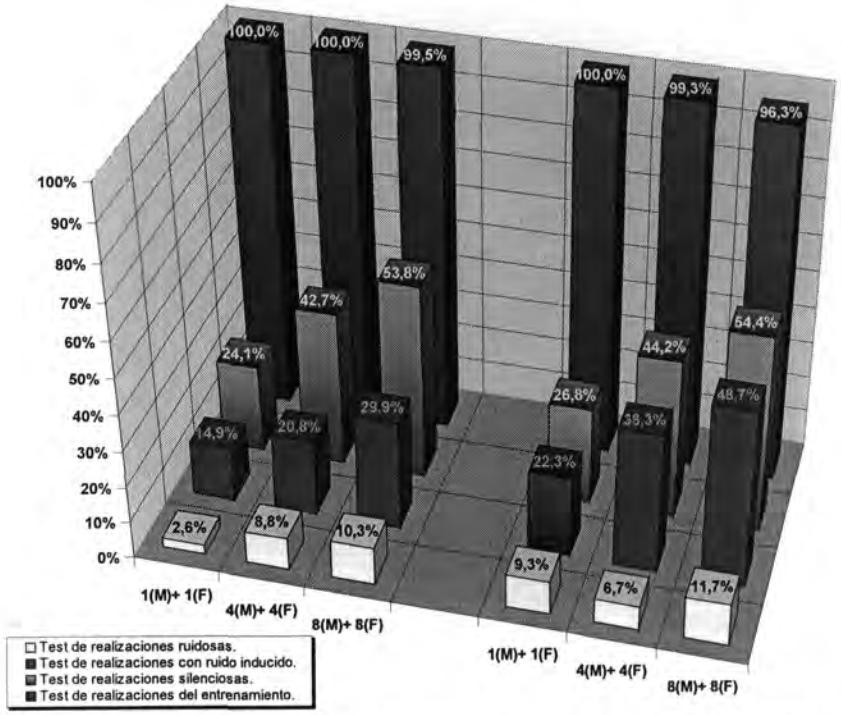

Figura C.34. Tasas de reconocimiento para el experimento FFT-$8_{(8)}$b (FFT-Cepstrum [4 coeficientes + 4 coeficientes delta <5 tramas>]. Cuantificación en 3 niveles: nivel 1° <16 índices> [1 coeficiente + 1 coeficiente delta], nivel 2° <48 índices> [1 coeficiente + 1 coeficiente delta], nivel 3° <192 índices> [2 coeficientes + 2 coeficientes delta]). La parte de la izquierda corresponde a los entrenamientos efectuados solamente con realizaciones silenciosas.

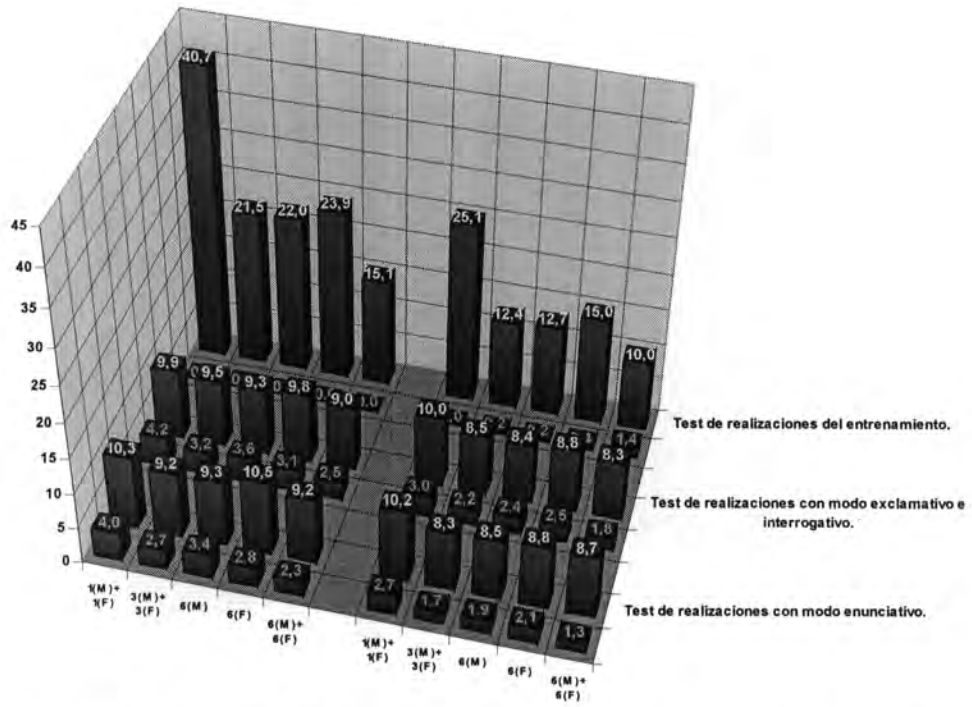

Figura C.35. Tasas de fiabilidad para el experimento FFT-$8_{(8)}$a (FFT-Cepstrum [4 coeficientes + 4 coeficientes delta <5 tramas>]. Cuantificación en 3 niveles: nivel 1º <16 índices> [1 coeficiente + 1 coeficiente delta], nivel 2º <48 índices> [1 coeficiente + 1 coeficiente delta], nivel 3º <192 índices> [2 coeficientes + 2 coeficientes delta]). En verde aparecen las tasas para los casos reconocidos correctamente. En rojo los casos clasificados erróneamente.

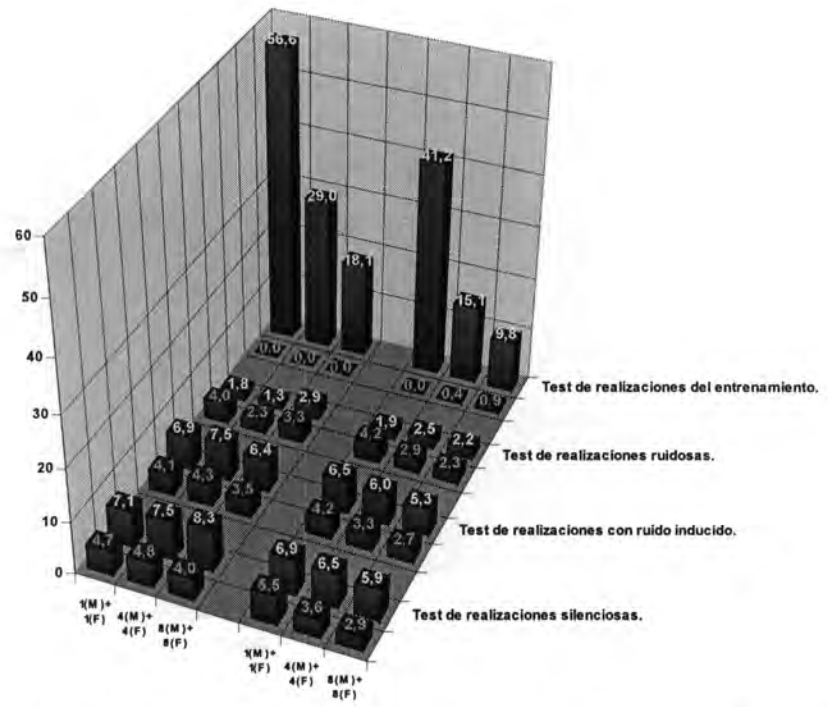

Figura C.36. Tasas de fiabilidad para el experimento FFT-$8_{(8)}$b (FFT-Cepstrum [4 coeficientes + 4 coeficientes delta <5 tramas>]. Cuantificación en 3 niveles: nivel 1º <16 índices> [1 coeficiente + 1 coeficiente delta], nivel 2º <48 índices> [1 coeficiente + 1 coeficiente delta], nivel 3º <192 índices> [2 coeficientes + 2 coeficientes delta]). En verde aparecen las tasas para los casos reconocidos correctamente. En rojo los casos clasificados erróneamente.

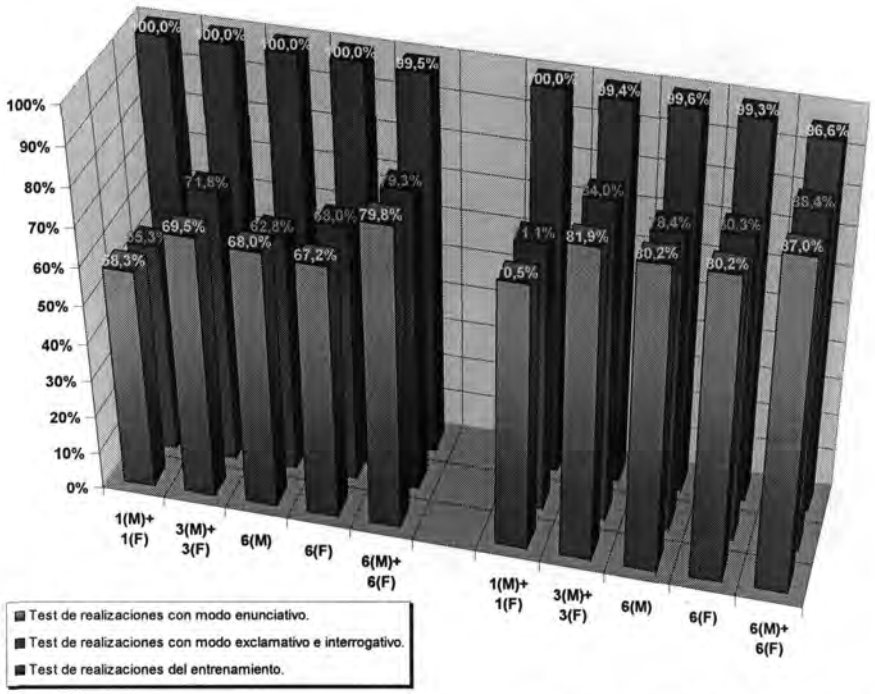

Figura C.37. Tasas de reconocimiento para el experimento FFT-$8_{(9)}$a (FFT-Cepstrum [4 coeficientes + 4 coeficientes delta <5 tramas>]. Cuantificación en 4 niveles: nivel 1° <16 índices> [1 coeficiente + 1 coeficiente delta], nivel 2° <48 índices> [1 coeficiente + 1 coeficiente delta], nivel 3° <64 índices> [2 coeficientes + 2 coeficientes delta], nivel 4° <128 índices> [1 coeficiente + 1 coeficiente delta]).

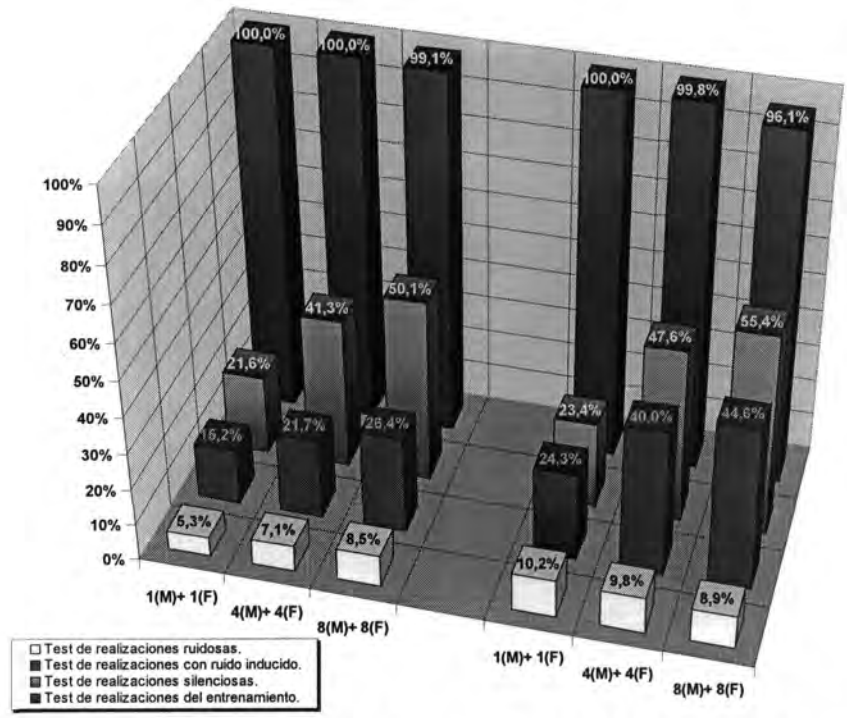

Figura C.38. Tasas de reconocimiento para el experimento FFT-$8_{(9)}$b (FFT-Cepstrum [4 coeficientes + 4 coeficientes delta <5 tramas>]. Cuantificación en 4 niveles: nivel 1° <16 índices> [1 coeficiente + 1 coeficiente delta], nivel 2° <48 índices> [1 coeficiente + 1 coeficiente delta], nivel 3° <64 índices> [2 coeficientes + 2 coeficientes delta], nivel 4° <128 índices> [1 coeficiente + 1 coeficiente delta]). La parte de la izquierda corresponde a los entrenamientos efectuados solamente con realizaciones silenciosas.

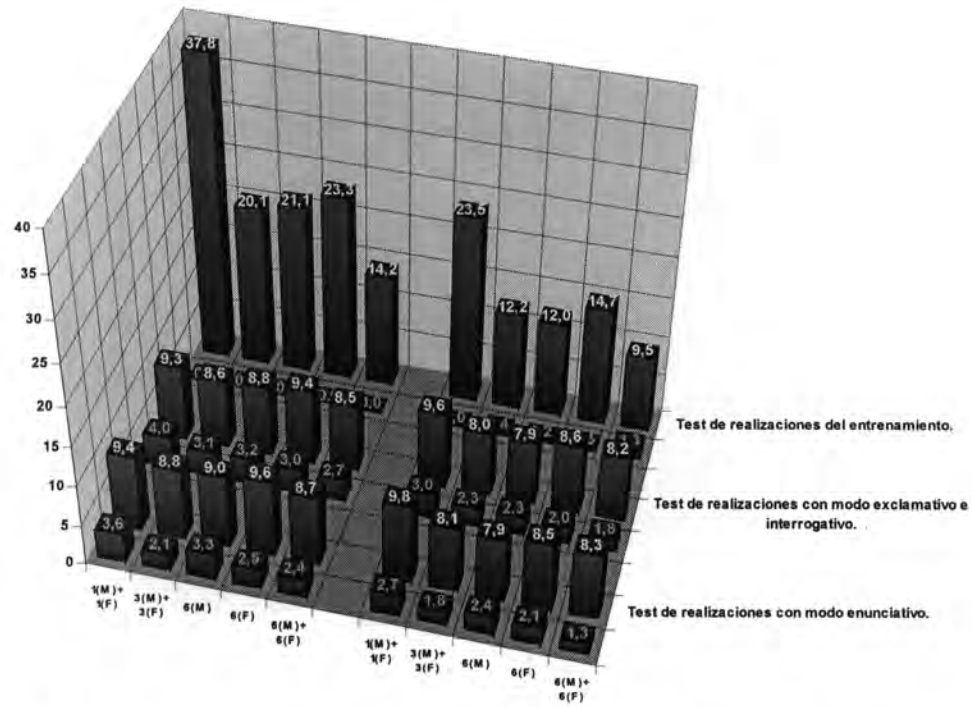

Figura C.39. Tasas de fiabilidad para el experimento FFT-8$_{(9)}$a (FFT-Cepstrum [4 coeficientes + 4 coeficientes delta <5 tramas>]. Cuantificación en 4 niveles: nivel 1º <16 índices> [1 coeficiente + 1 coeficiente delta], nivel 2º <48 índices> [1 coeficiente + 1 coeficiente delta], nivel 3º <64 índices> [2 coeficientes + 2 coeficientes delta], nivel 4º <128 índices> [1 coeficiente + 1 coeficiente delta]). En verde aparecen las tasas para los casos reconocidos correctamente. En rojo los casos clasificados erróneamente.

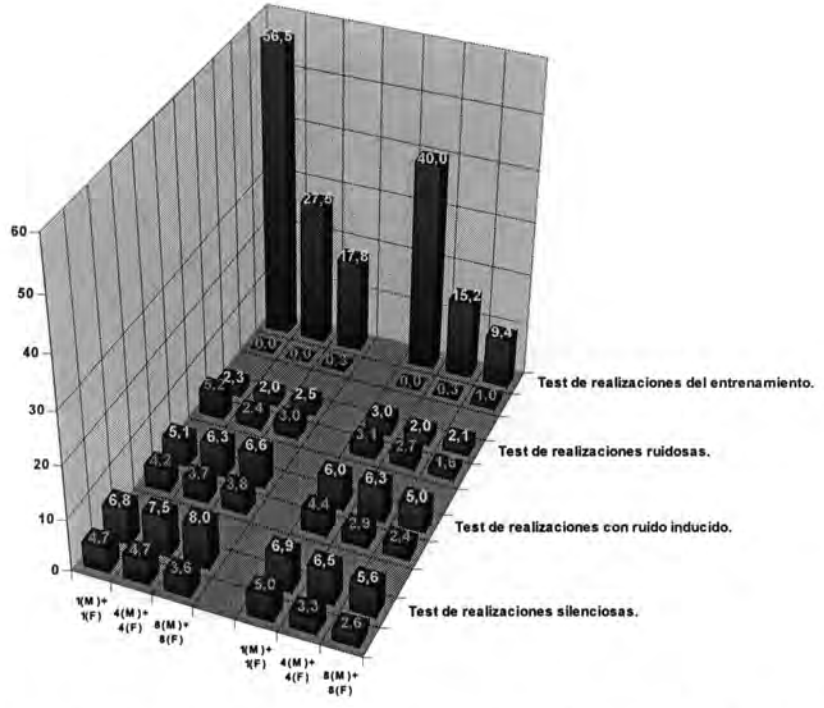

Figura C.40. Tasas de fiabilidad para el experimento FFT-8$_{(9)}$b (FFT-Cepstrum [4 coeficientes + 4 coeficientes delta <5 tramas>]. Cuantificación en 4 niveles: nivel 1º <16 índices> [1 coeficiente + 1 coeficiente delta], nivel 2º <48 índices> [1 coeficiente + 1 coeficiente delta], nivel 3º <64 índices> [2 coeficientes + 2 coeficientes delta], nivel 4º <128 índices> [1 coeficiente + 1 coeficiente delta]). En verde aparecen las tasas para los casos reconocidos correctamente. En rojo los casos clasificados erróneamente.

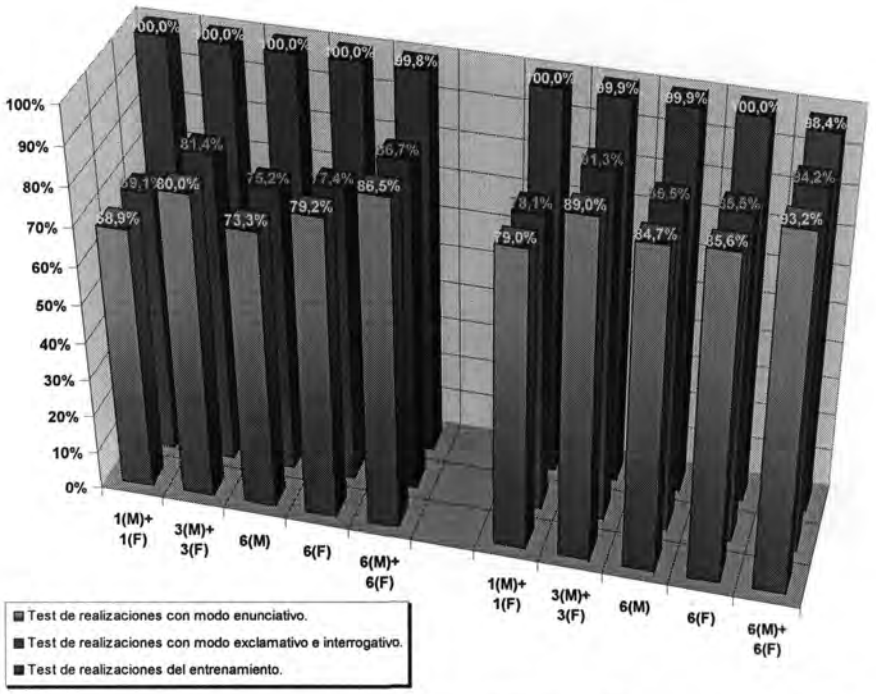

Figura C.41. Tasas de reconocimiento para el experimento FFT-$8_{(10)}$a (Tono fundamental + FFT-Cepstrum [4 coeficientes + 4 coeficientes delta <5 tramas>]). La parte de la izquierda corresponde a los entrenamientos efectuados solamente con realizaciones con modo enunciativo.

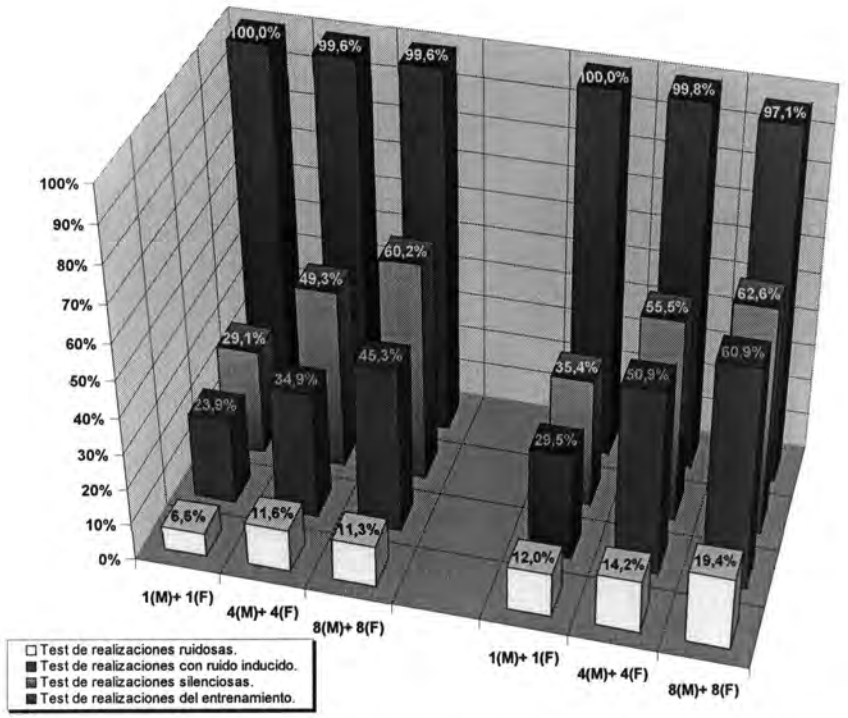

Figura C.42. Tasas de reconocimiento para el experimento FFT-$8_{(10)}$b (Tono fundamental + FFT-Cepstrum [4 coeficientes + 4 coeficientes delta <5 tramas>]). La parte de la izquierda corresponde a los entrenamientos efectuados solamente con realizaciones silenciosas.

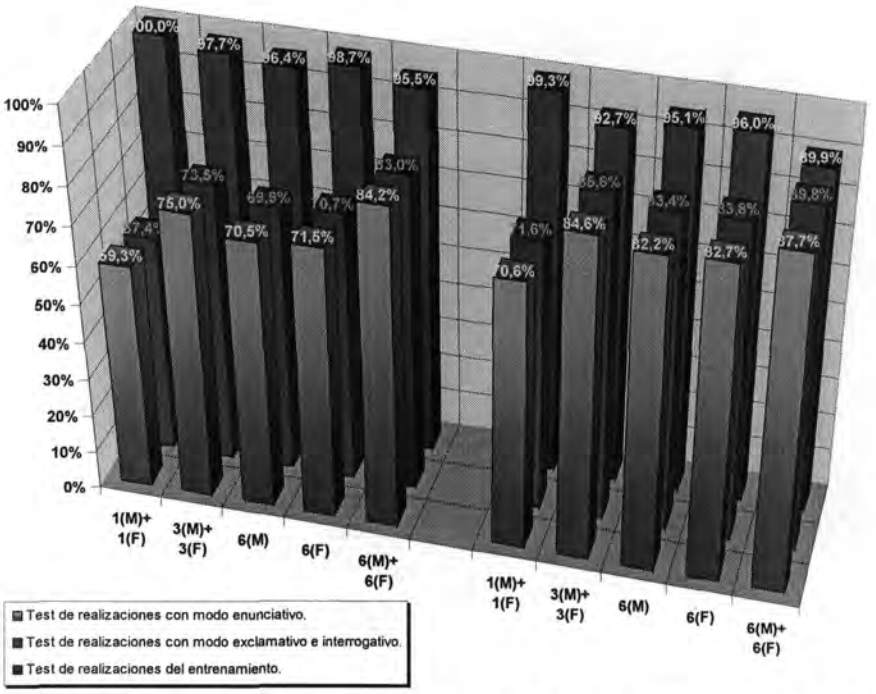

Figura C.45. Tasas de reconocimiento para el experimento FFT-$8_{(11)}$a (Tono fundamental + FFT-Cepstrum [4 coeficientes + 4 coeficientes delta <5 tramas>]. Cuantificación en 2 niveles: nivel 1° <2 índices> [Tono fundamental], nivel 2° <254 índices> [4 coeficientes + 4 coeficientes delta]). La parte de la izquierda corresponde a los entrenamientos efectuados solamente con realizaciones con modo enunciativo.

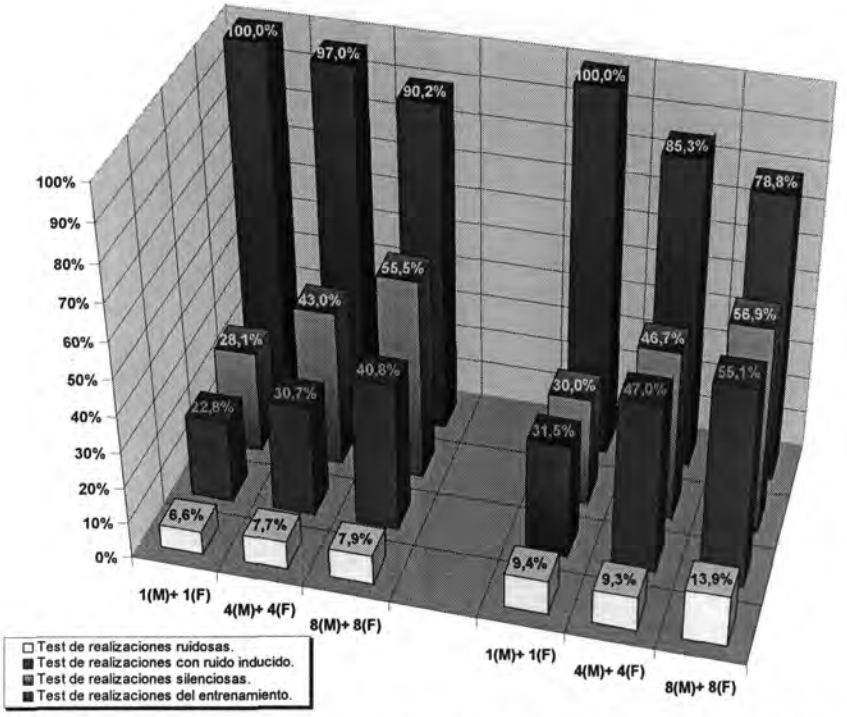

Figura C.46. Tasas de reconocimiento para el experimento FFT-$8_{(11)}$b (Tono fundamental + FFT-Cepstrum [4 coeficientes + 4 coeficientes delta <5 tramas>]. Cuantificación en 2 niveles: nivel 1° <2 índices> [Tono fundamental], nivel 2° <254 índices> [4 coeficientes + 4 coeficientes delta]). La parte de la izquierda corresponde a los entrenamientos efectuados solamente con realizaciones silenciosas.

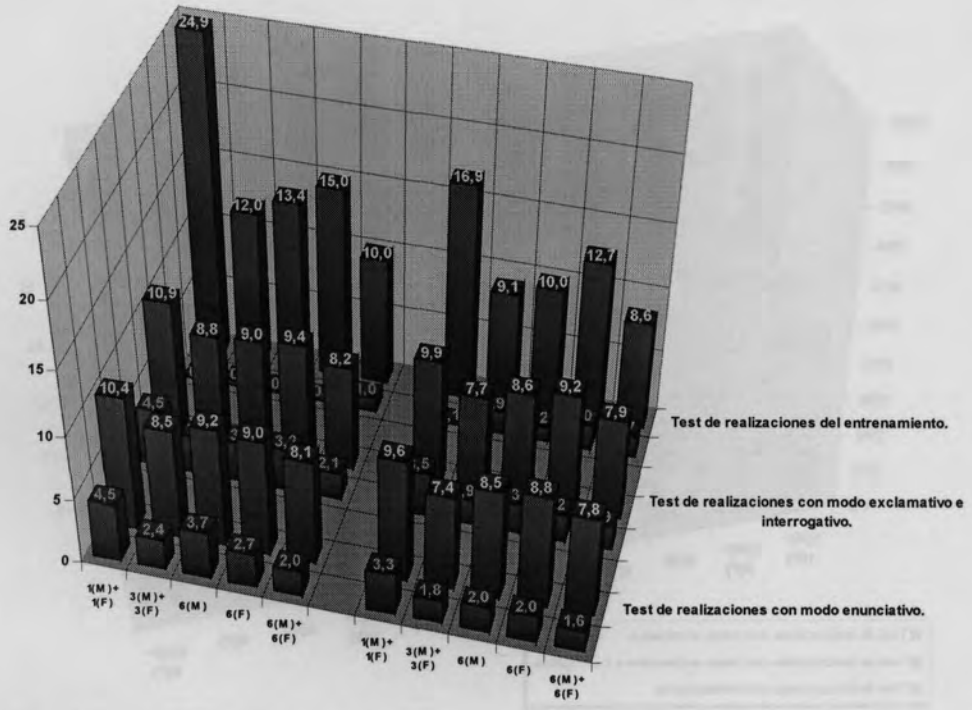

Figura C.47. Tasas de fiabilidad para el experimento FFT-$8_{(11)}$a (Tono fundamental + FFT-Cepstrum [4 coeficientes + 4 coeficientes delta <5 tramas>]. Cuantificación en 2 niveles: nivel 1º <2 índices> [Tono fundamental], nivel 2º <254 índices> [4 coeficientes + 4 coeficientes delta]). En verde aparecen las tasas para los casos reconocidos correctamente. En rojo los casos clasificados erróneamente.

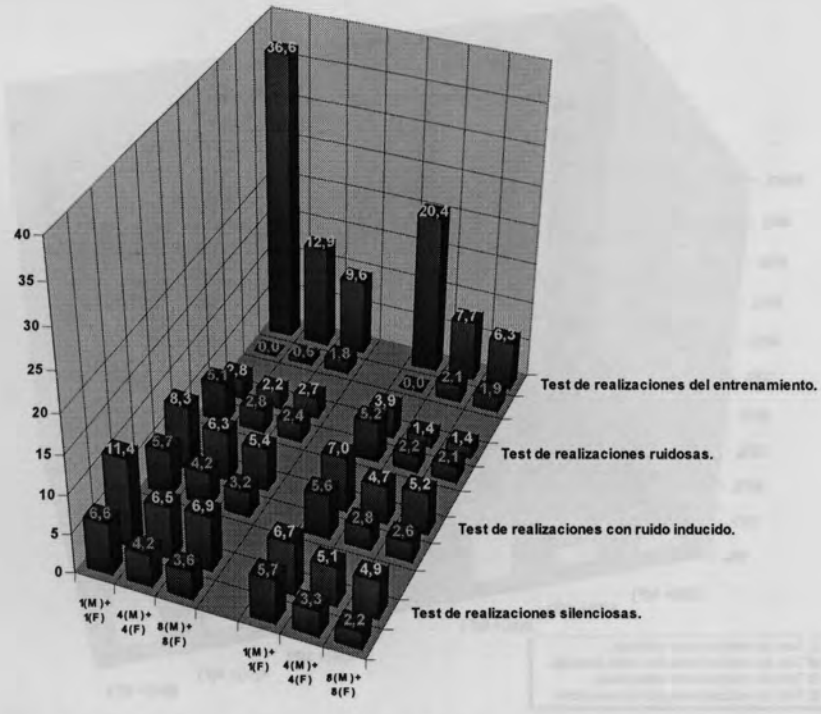

Figura C.48. Tasas de fiabilidad para el experimento FFT-$8_{(11)}$b (Tono fundamental + FFT-Cepstrum [4 coeficientes + 4 coeficientes delta <5 tramas>]. Cuantificación en 2 niveles: nivel 1º <2 índices> [Tono fundamental], nivel 2º <254 índices> [4 coeficientes + 4 coeficientes delta]). En verde aparecen las tasas para los casos reconocidos correctamente. En rojo los casos clasificados erróneamente.

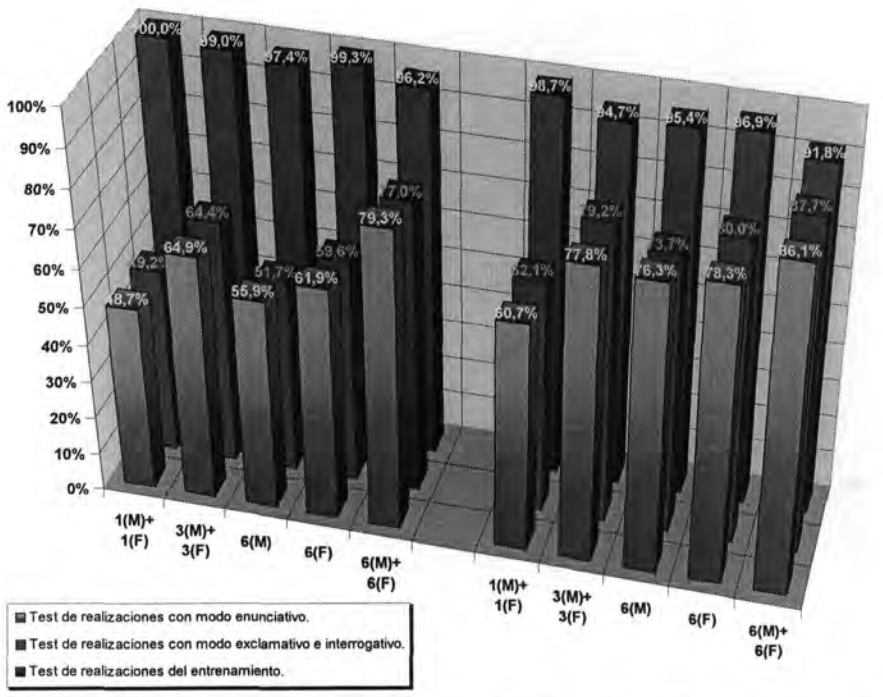

Figura C.49. Tasas de reconocimiento para el experimento FFT-$8_{(12)}$a (Tono fundamental + FFT-Cepstrum [4 coeficientes + 4 coeficientes delta <5 tramas>]. Cuantificación en 2 niveles: nivel 1º <4 índices> [Tono fundamental], nivel 2º <252 índices> [4 coeficientes + 4 coeficientes delta]). La parte de la izquierda corresponde a los entrenamientos efectuados solamente con realizaciones con modo enunciativo.

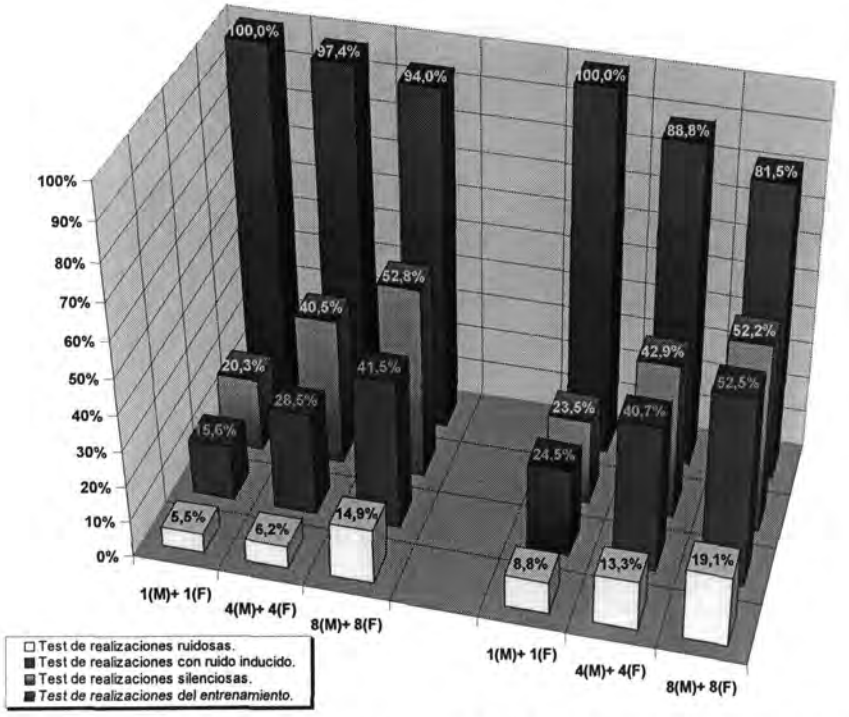

Figura C.50. Tasas de reconocimiento para el experimento FFT-$8_{(12)}$b (Tono fundamental + FFT-Cepstrum [4 coeficientes + 4 coeficientes delta <5 tramas>]. Cuantificación en 2 niveles: nivel 1º <4 índices> [Tono fundamental], nivel 2º <252 índices> [4 coeficientes + 4 coeficientes delta]). La parte de la izquierda corresponde a los entrenamientos efectuados solamente con realizaciones silenciosas.

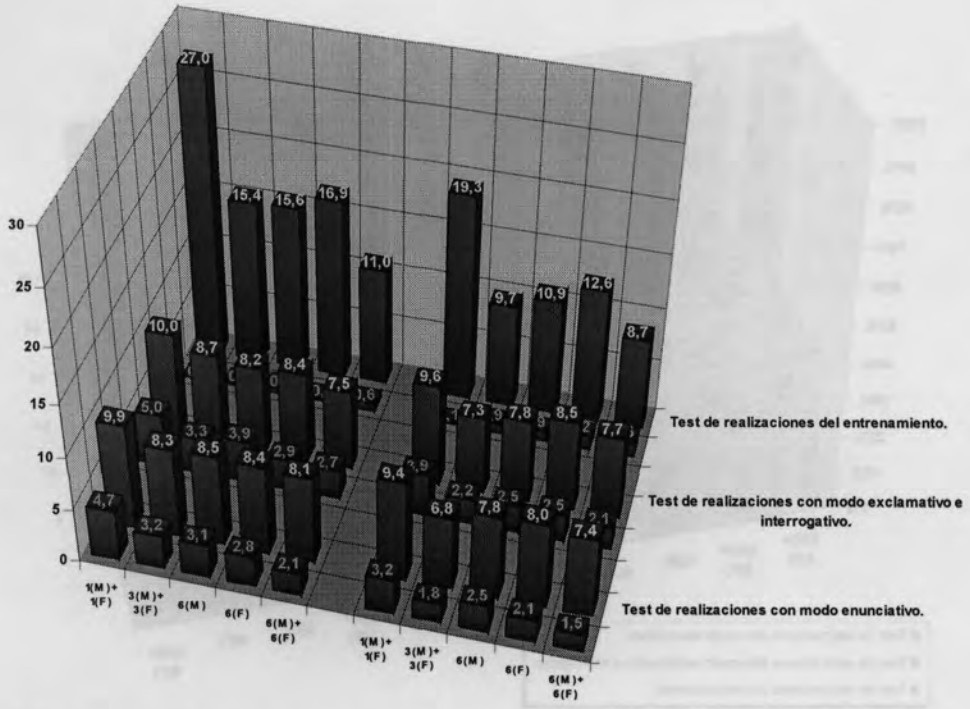

Figura C.51. Tasas de fiabilidad para el experimento FFT-$8_{(12)}$a (Tono fundamental + FFT-Cepstrum [4 coeficientes + 4 coeficientes delta <5 tramas>]. Cuantificación en 2 niveles: nivel 1º <4 índices> [Tono fundamental], nivel 2º <252 índices> [4 coeficientes + 4 coeficientes delta]). En verde aparecen las tasas para los casos reconocidos correctamente. En rojo los casos clasificados erróneamente.

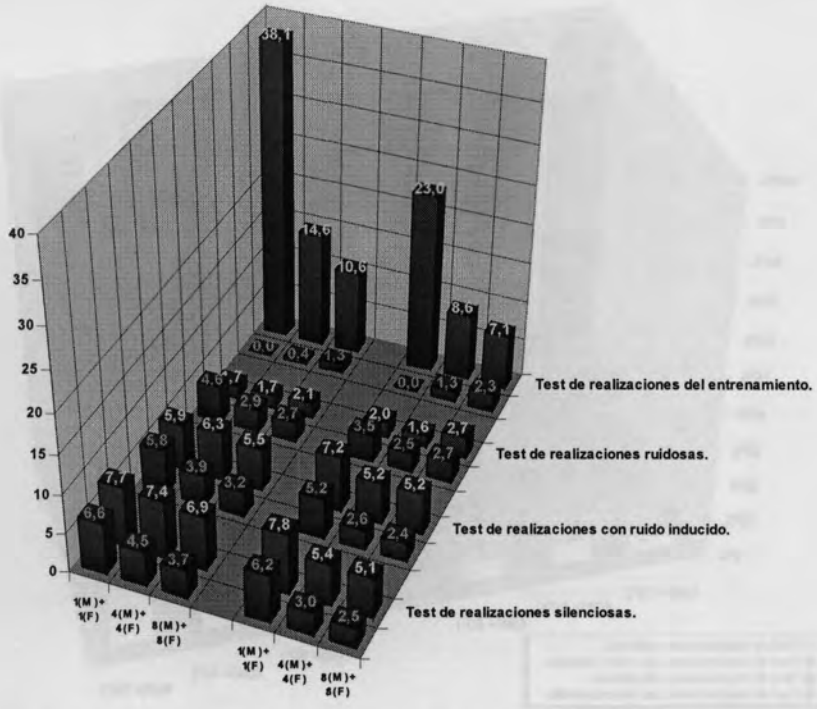

Figura C.52. Tasas de fiabilidad para el experimento FFT-$8_{(12)}$b (Tono fundamental + FFT-Cepstrum [4 coeficientes + 4 coeficientes delta <5 tramas>]. Cuantificación en 2 niveles: nivel 1º <4 índices> [Tono fundamental], nivel 2º <252 índices> [4 coeficientes + 4 coeficientes delta]). En verde aparecen las tasas para los casos reconocidos correctamente. En rojo los casos clasificados erróneamente.

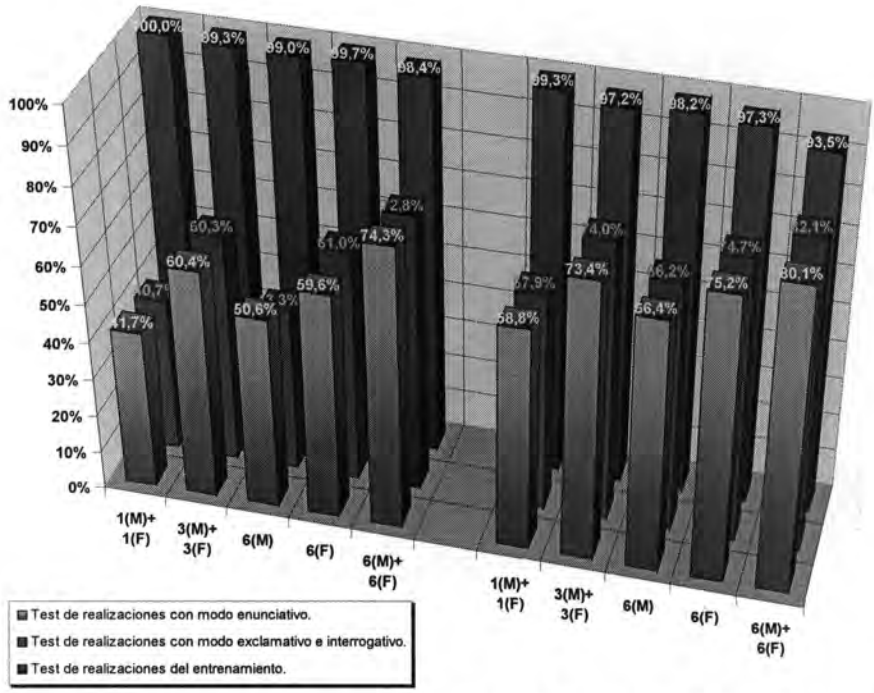

Figura C.53. Tasas de reconocimiento para el experimento FFT-$8_{(13)}$a (Tono fundamental + FFT-Cepstrum [4 coeficientes + 4 coeficientes delta <5 tramas>]. Cuantificación en 2 niveles: nivel 1º <8 índices> [Tono fundamental], nivel 2º <248 índices> [4 coeficientes + 4 coeficientes delta]). La parte de la izquierda corresponde a los entrenamientos efectuados solamente con realizaciones con modo enunciativo.

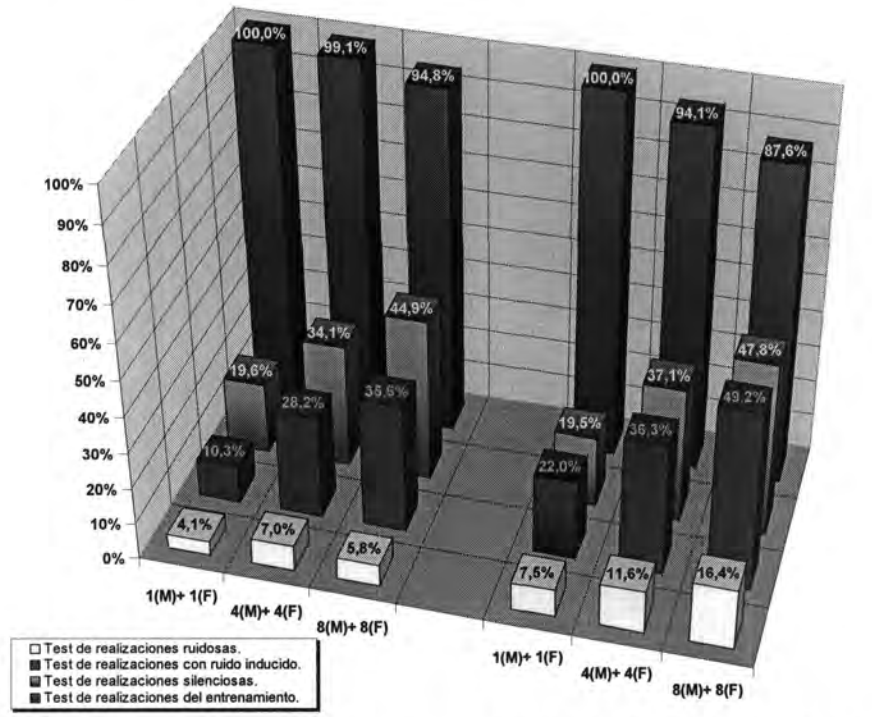

Figura C.54. Tasas de reconocimiento para el experimento FFT-$8_{(13)}$b (Tono fundamental + FFT-Cepstrum [4 coeficientes + 4 coeficientes delta <5 tramas>]. Cuantificación en 2 niveles: nivel 1º <8 índices> [Tono fundamental], nivel 2º <248 índices> [4 coeficientes + 4 coeficientes delta]). La parte de la izquierda corresponde a los entrenamientos efectuados solamente con realizaciones silenciosas.

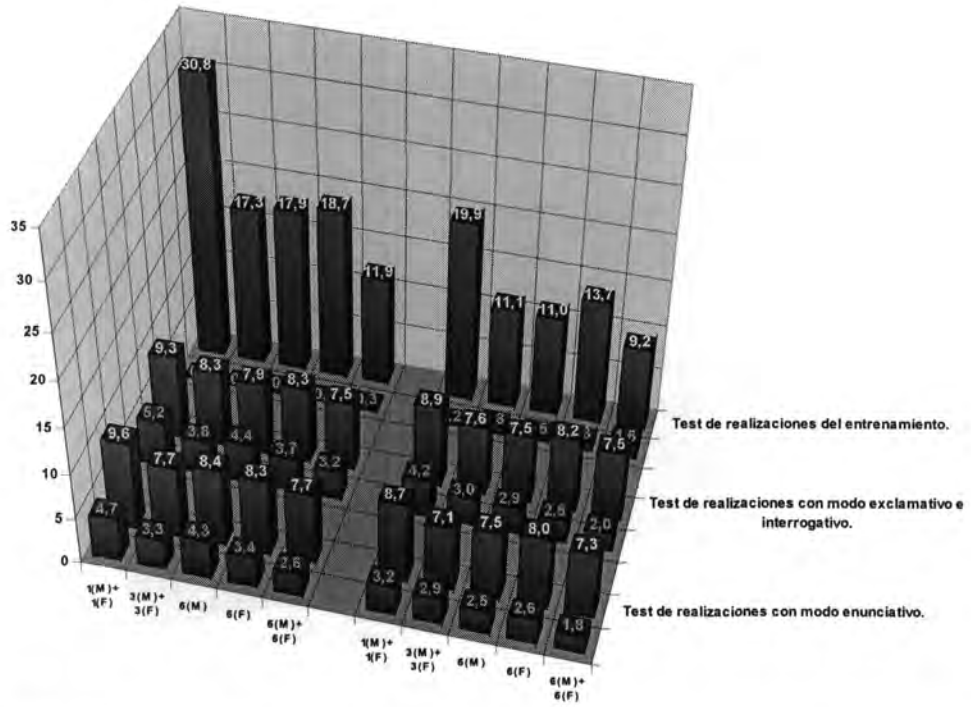

Figura C.55. Tasas de fiabilidad para el experimento FFT-$8_{(13)}$a (Tono fundamental + FFT-Cepstrum [4 coeficientes + 4 coeficientes delta <5 tramas>]. Cuantificación en 2 niveles: nivel 1º <8 índices> [Tono fundamental], nivel 2º <248 índices> [4 coeficientes + 4 coeficientes delta]). En verde aparecen las tasas para los casos reconocidos correctamente. En rojo los casos clasificados erróneamente.

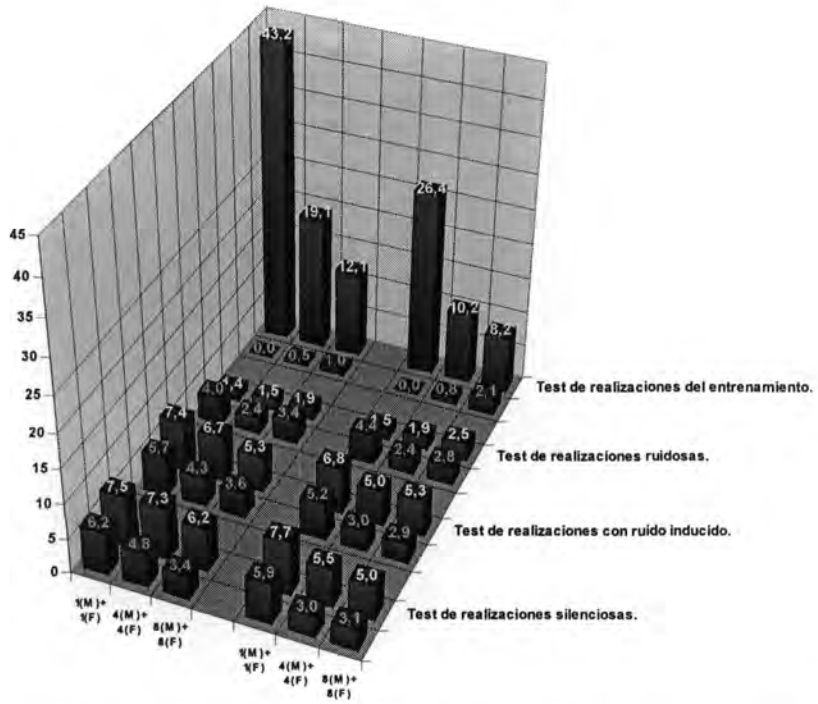

Figura C.56. Tasas de fiabilidad para el experimento FFT-$8_{(13)}$b (Tono fundamental + FFT-Cepstrum [4 coeficientes + 4 coeficientes delta <5 tramas>]. Cuantificación en 2 niveles: nivel 1º <8 índices> [Tono fundamental], nivel 2º <248 índices> [4 coeficientes + 4 coeficientes delta]). En verde aparecen las tasas para los casos reconocidos correctamente. En rojo los casos clasificados erróneamente.

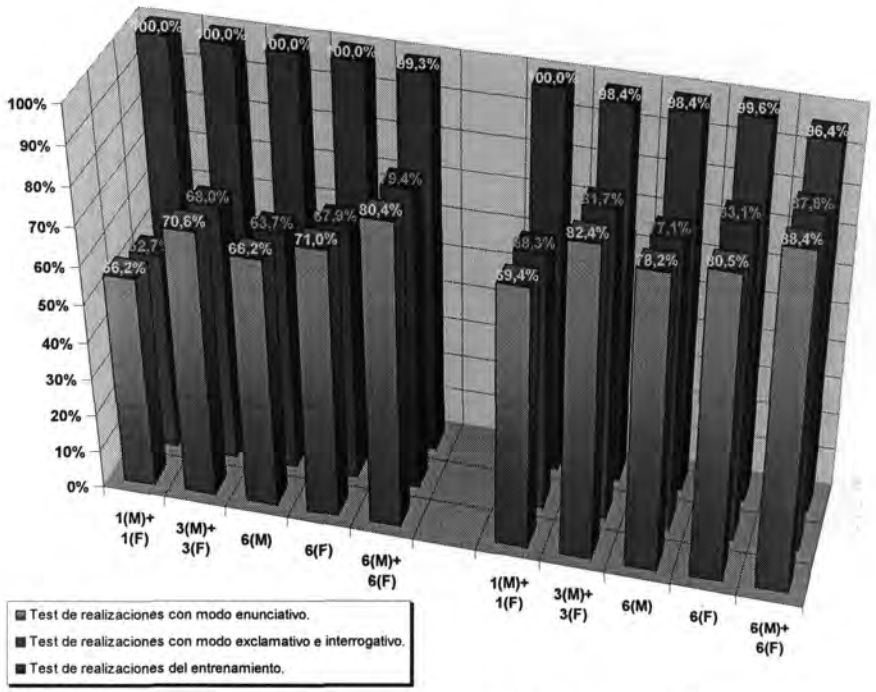

Figura C.57. Tasas de reconocimiento para el experimento FFT-$8_{(14)}$a (Tono fundamental + FFT-Cepstrum [4 coeficientes + 4 coeficientes delta <5 tramas>]. Cuantificación en 3 niveles: nivel 1º <2 índices> [Tono fundamental], nivel 2º <62 índices> [2 coeficiente], nivel 3º <192 índices> [2 coeficientes + 4 coeficientes delta]). La parte de la izquierda corresponde a los entrenamientos efectuados solamente con realizaciones con modo enunciativo.

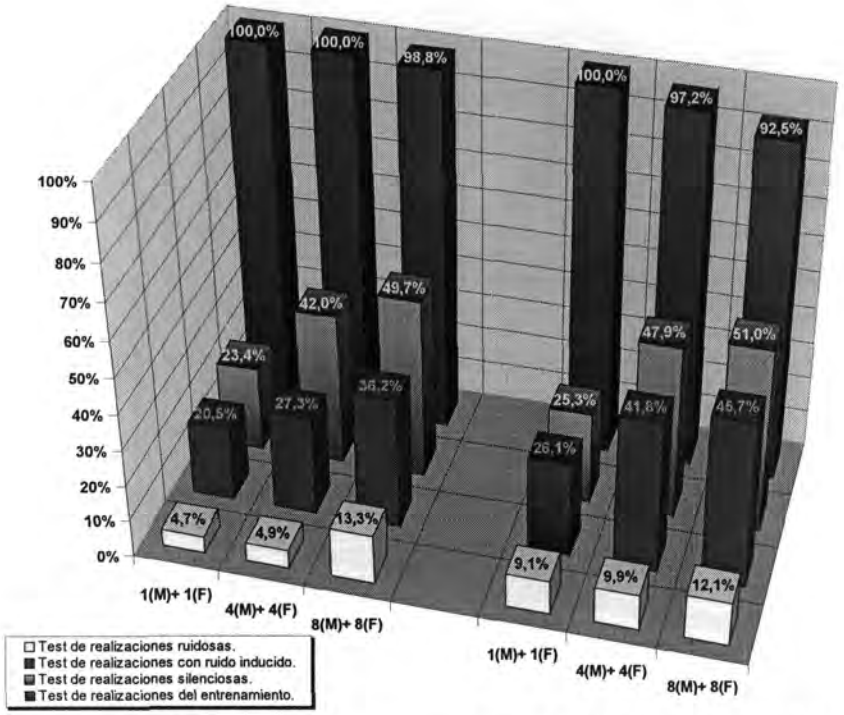

Figura C.58. Tasas de reconocimiento para el experimento FFT-$8_{(14)}$b (Tono fundamental + FFT-Cepstrum [4 coeficientes + 4 coeficientes delta <5 tramas>]. Cuantificación en 3 niveles: nivel 1º <2 índices> [Tono fundamental], nivel 2º <62 índices> [2 coeficiente], nivel 3º <192 índices> [2 coeficientes + 4 coeficientes delta]). La parte de la izquierda corresponde a los entrenamientos efectuados solamente con realizaciones silenciosas.

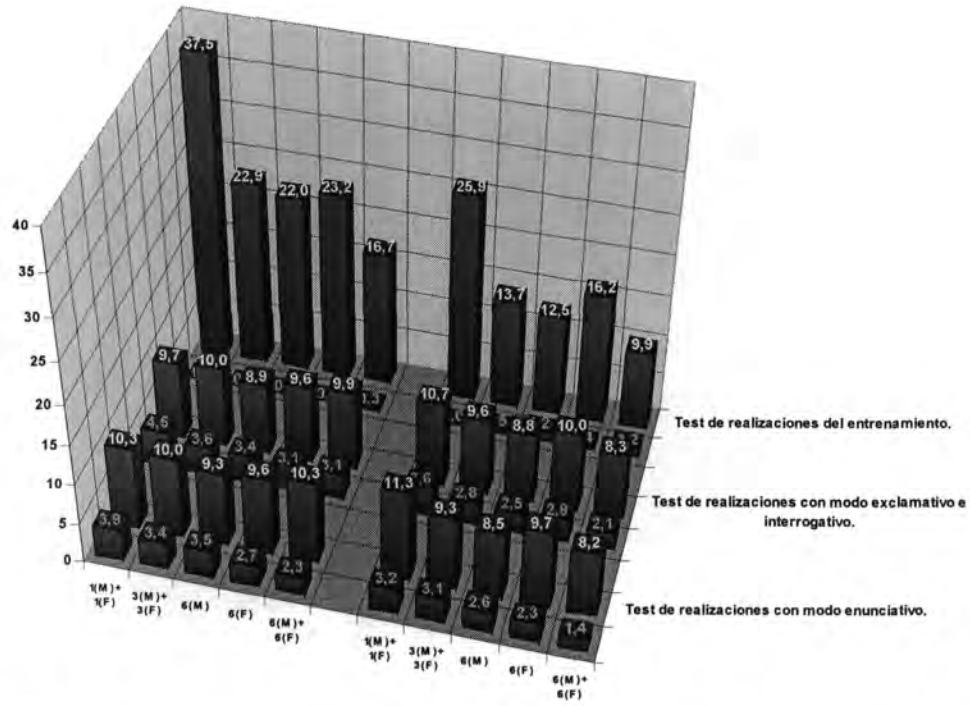

Figura C.59. Tasas de fiabilidad para el experimento FFT-$8_{(14)}$a (Tono fundamental + FFT-Cepstrum [4 coeficientes + 4 coeficientes delta <5 tramas>]. Cuantificación en 3 niveles: nivel 1º <2 índices> [Tono fundamental], nivel 2º <62 índices> [2 coeficiente], nivel 3º <192 índices> [2 coeficientes + 4 coeficientes delta]). En verde aparecen las tasas para los casos reconocidos correctamente. En rojo los casos clasificados erróneamente.

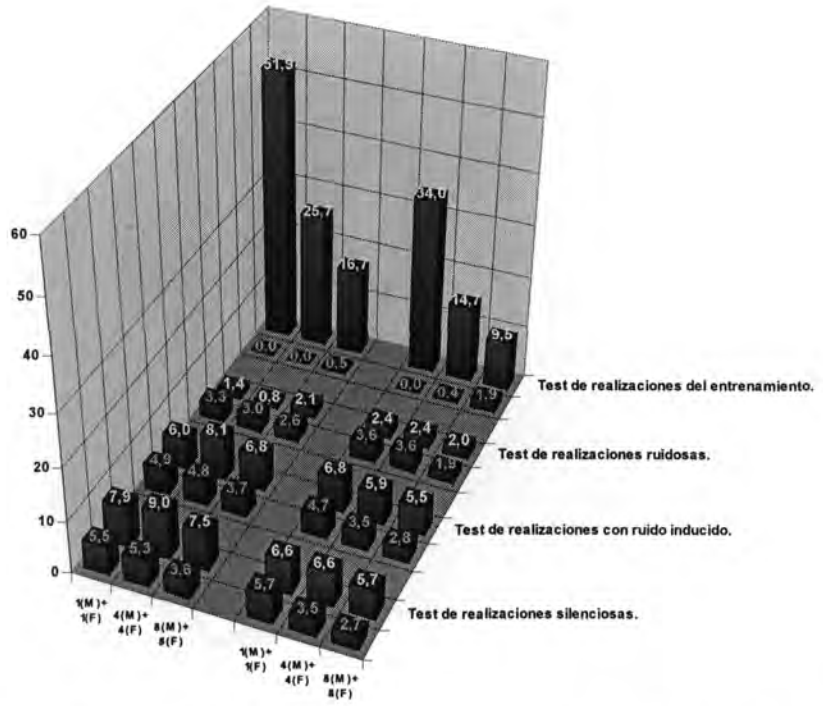

Figura C.60. Tasas de fiabilidad para el experimento FFT-$8_{(14)}$b (Tono fundamental + FFT-Cepstrum [4 coeficientes + 4 coeficientes delta <5 tramas>]. Cuantificación en 3 niveles: nivel 1º <2 índices> [Tono fundamental], nivel 2º <62 índices> [2 coeficiente], nivel 3º <192 índices> [2 coeficientes + 4 coeficientes delta]). En verde aparecen las tasas para los casos reconocidos correctamente. En rojo los casos clasificados erróneamente.

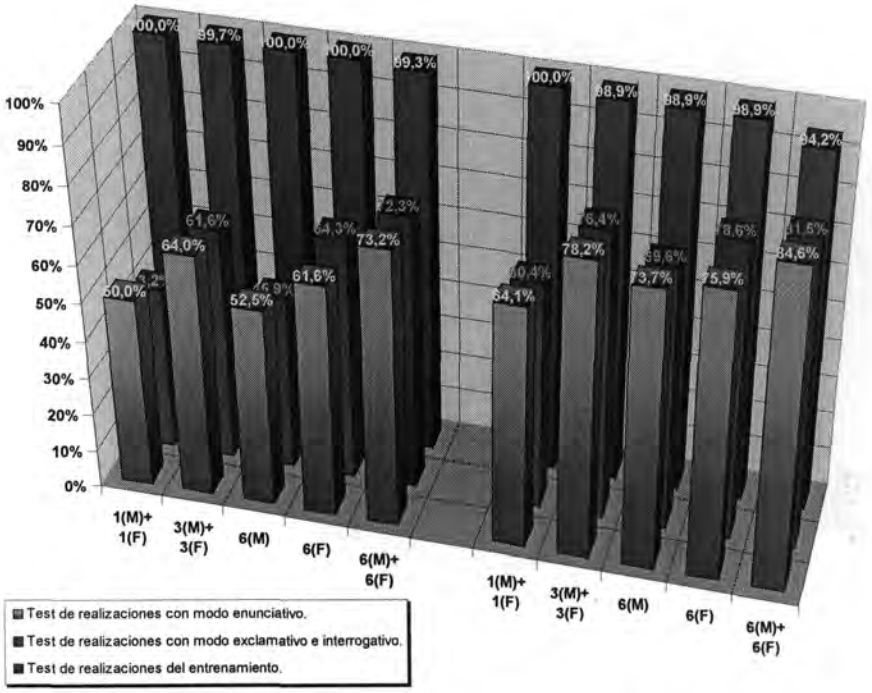

Figura C.61. Tasas de reconocimiento para el experimento FFT-$8_{(15)}$a (Tono fundamental + FFT-Cepstrum [4 coeficientes + 4 coeficientes delta <5 tramas>]. Cuantificación en 3 niveles: nivel 1º <4 índices> [Tono fundamental], nivel 2º <60 índices> [2 coeficiente], nivel 3º <192 índices> [2 coeficientes + 4 coeficientes delta]). La parte de la izquierda corresponde a los entrenamientos efectuados solamente con realizaciones con modo enunciativo.

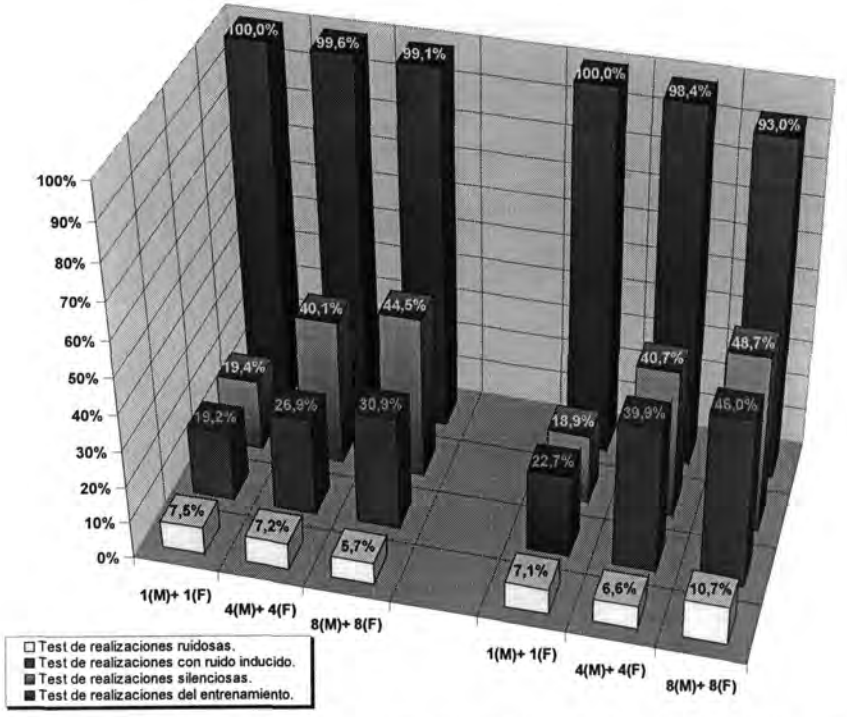

Figura C.62. Tasas de reconocimiento para el experimento FFT-$8_{(15)}$b (Tono fundamental + FFT-Cepstrum [4 coeficientes + 4 coeficientes delta <5 tramas>]. Cuantificación en 3 niveles: nivel 1º <4 índices> [Tono fundamental], nivel 2º <60 índices> [2 coeficiente], nivel 3º <192 índices> [2 coeficientes + 4 coeficientes delta]). La parte de la izquierda corresponde a los entrenamientos efectuados solamente con realizaciones silenciosas.

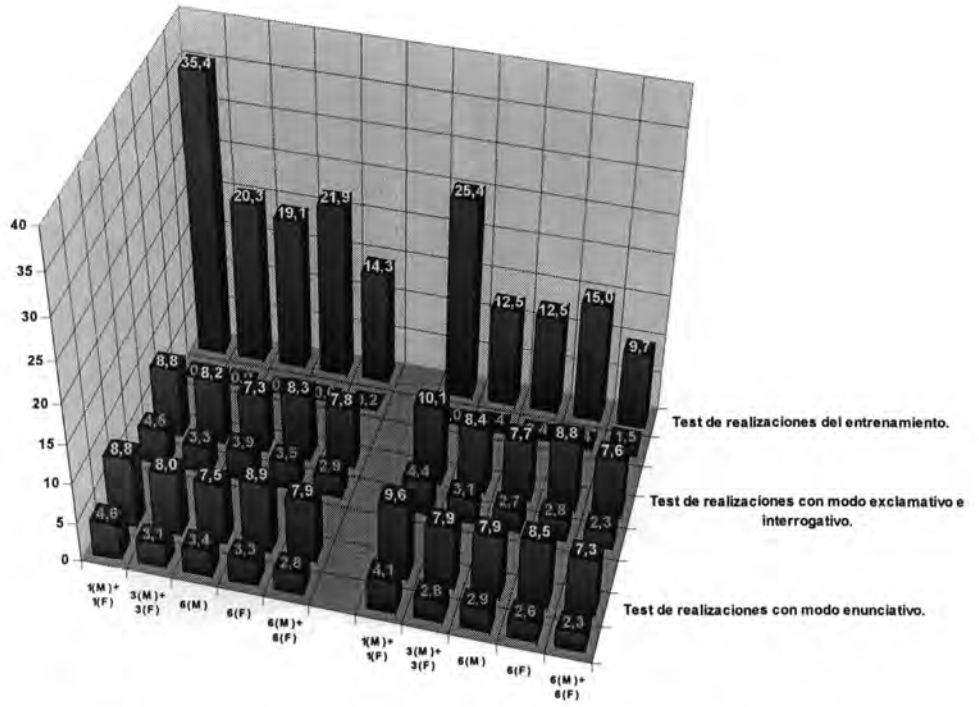

Figura C.63. Tasas de fiabilidad para el experimento FFT-$8_{(15)}$a (Tono fundamental + FFT-Cepstrum [4 coeficientes + 4 coeficientes delta <5 tramas>]. Cuantificación en 3 niveles: nivel 1° <4 índices> [Tono fundamental], nivel 2° <60 índices> [2 coeficiente], nivel 3° <192 índices> [2 coeficientes + 4 coeficientes delta]). En verde aparecen las tasas para los casos reconocidos correctamente. En rojo los casos clasificados erróneamente.

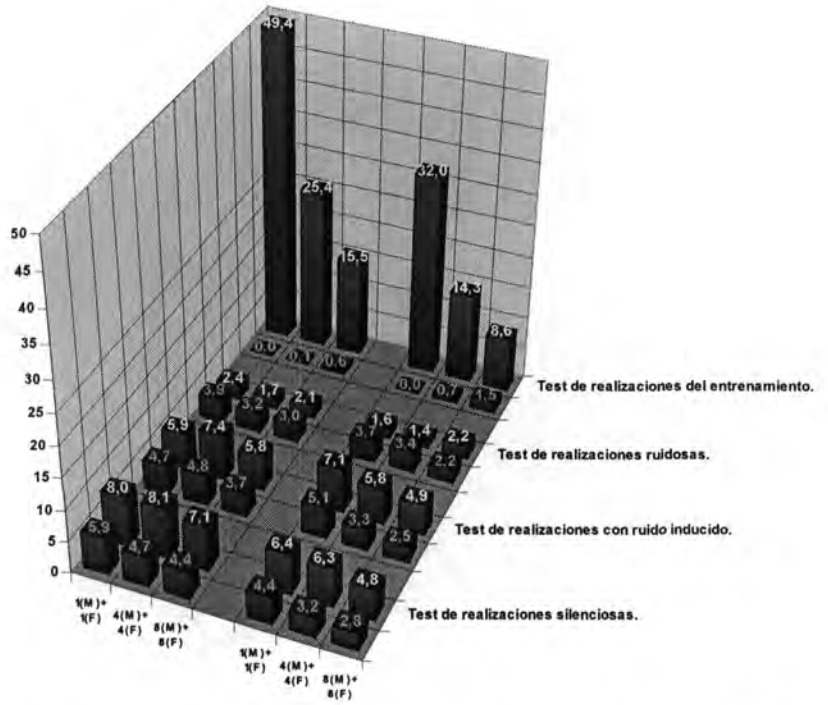

Figura C.64. Tasas de fiabilidad para el experimento FFT-$8_{(15)}$b (Tono fundamental + FFT-Cepstrum [4 coeficientes + 4 coeficientes delta <5 tramas>]. Cuantificación en 3 niveles: nivel 1° <4 índices> [Tono fundamental], nivel 2° <60 índices> [2 coeficiente], nivel 3° <192 índices> [2 coeficientes + 4 coeficientes delta]). En verde aparecen las tasas para los casos reconocidos correctamente. En rojo los casos clasificados erróneamente.

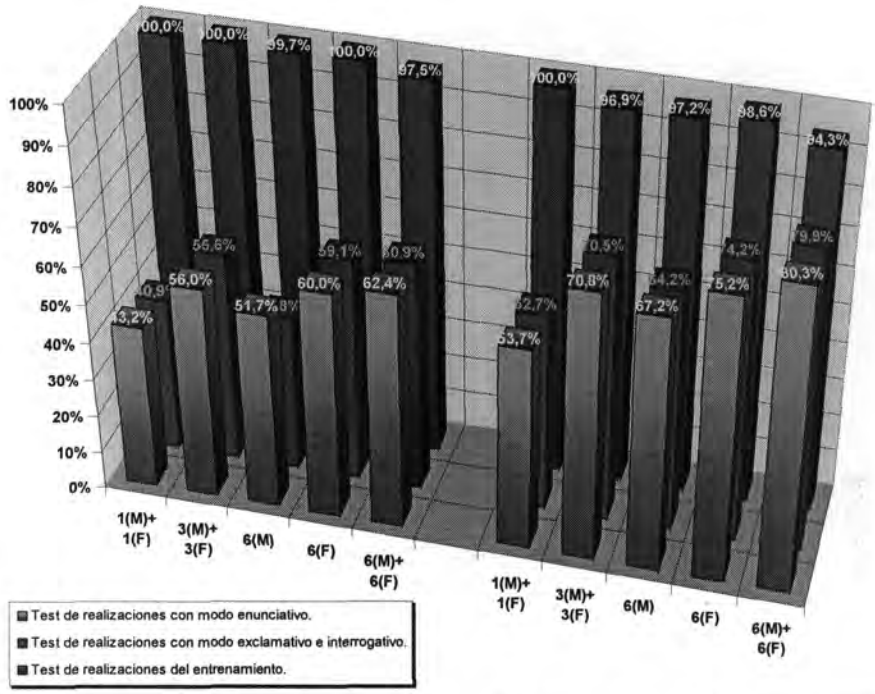

Figura C.65. Tasas de reconocimiento para el experimento FFT-$8_{(16)}$a (Tono fundamental + FFT-Cepstrum [4 coeficientes + 4 coeficientes delta <5 tramas>]. Cuantificación en 3 niveles: nivel 1º <8 índices> [Tono fundamental], nivel 2º <56 índices> [2 coeficiente], nivel 3º <192 índices> [2 coeficientes + 4 coeficientes delta]). La parte de la izquierda corresponde a los entrenamientos efectuados solamente con realizaciones con modo enunciativo.

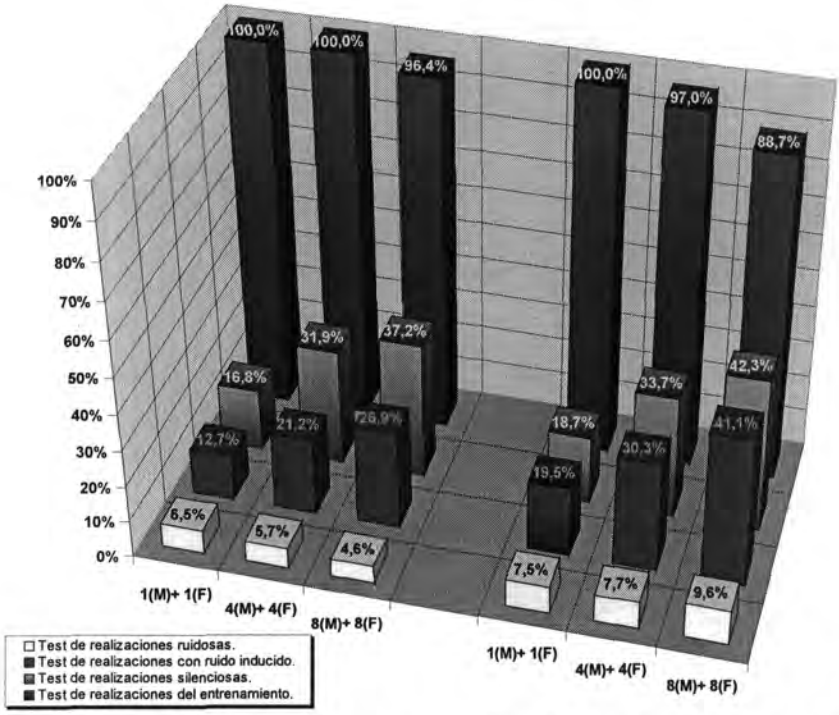

Figura C.66. Tasas de reconocimiento para el experimento FFT-$8_{(16)}$b (Tono fundamental + FFT-Cepstrum [4 coeficientes + 4 coeficientes delta <5 tramas>]. Cuantificación en 3 niveles: nivel 1º <8 índices> [Tono fundamental], nivel 2º <56 índices> [2 coeficiente], nivel 3º <192 índices> [2 coeficientes + 4 coeficientes delta]). La parte de la izquierda corresponde a los entrenamientos efectuados solamente con realizaciones silenciosas.

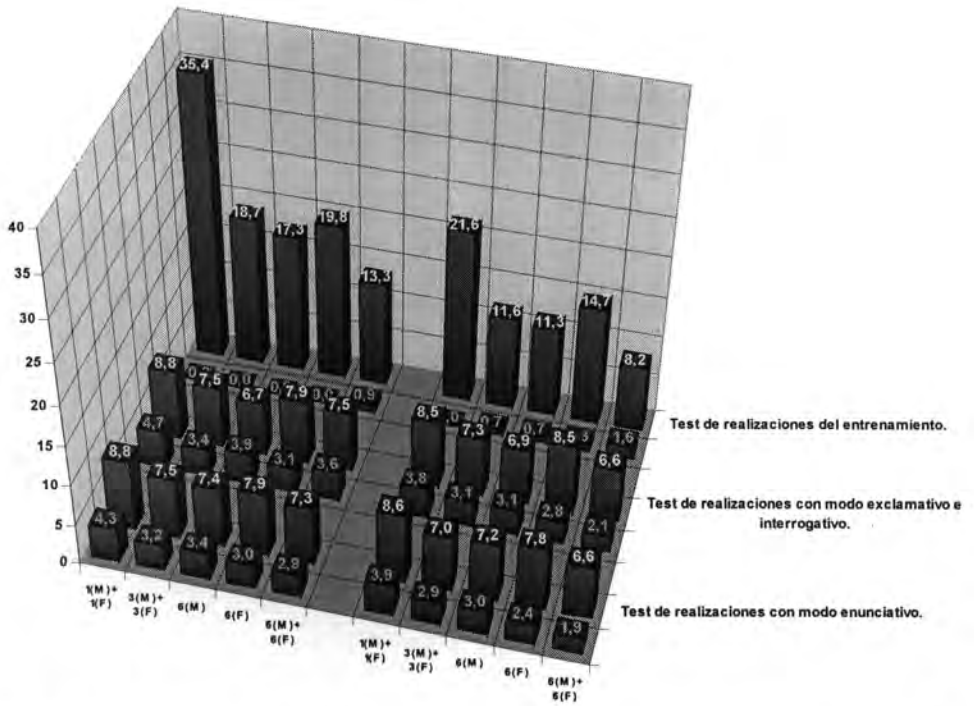

Figura C.67. Tasas de fiabilidad para el experimento FFT-$8_{(16)}$a (Tono fundamental + FFT-Cepstrum [4 coeficientes + 4 coeficientes delta <5 tramas>]. Cuantificación en 3 niveles: nivel 1º <8 índices> [Tono fundamental], nivel 2º <56 índices> [2 coeficiente], nivel 3º <192 índices> [2 coeficientes + 4 coeficientes delta]). En verde aparecen las tasas para los casos reconocidos correctamente. En rojo los casos clasificados erróneamente.

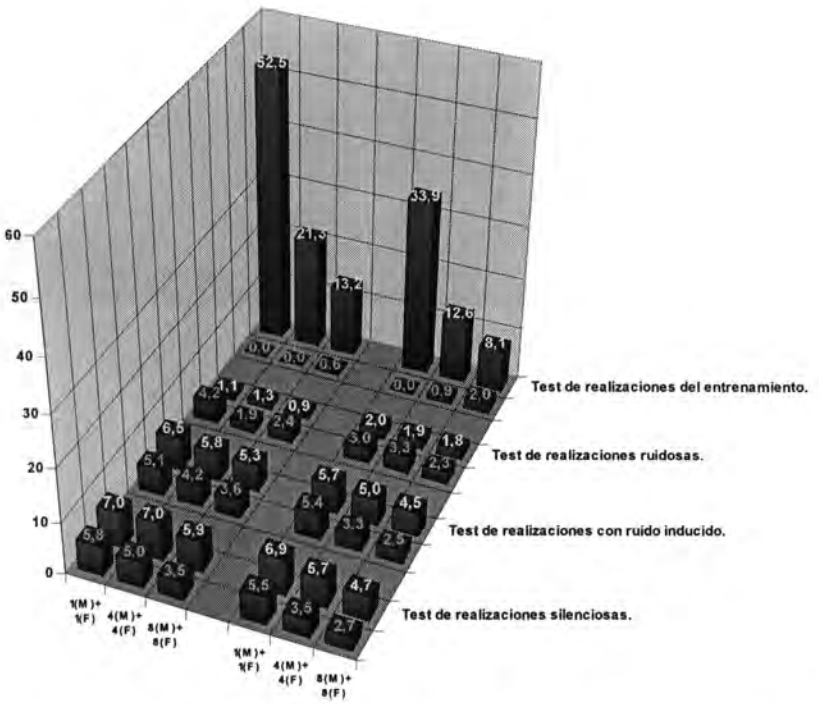

Figura C.68. Tasas de fiabilidad para el experimento FFT-$8_{(16)}$b (Tono fundamental + FFT-Cepstrum [4 coeficientes + 4 coeficientes delta <5 tramas>]. Cuantificación en 3 niveles: nivel 1º <8 índices> [Tono fundamental], nivel 2º <56 índices> [2 coeficiente], nivel 3º <192 índices> [2 coeficientes + 4 coeficientes delta]). En verde aparecen las tasas para los casos reconocidos correctamente. En rojo los casos clasificados erróneamente.

APÉNDICE D. DETALLES DE IMPLEMENTACIÓN DEL RECONOCEDOR DE VOZ DE PALABRAS AISLADAS EN EL TMS320C31-60

D.1. ARQUITECTURA DEL PROCESADOR DE SEÑAL DIGITAL TMS320C31

El TMS320C31 [TEX94] es un procesador digital de señal con coma flotante de 32 bits, que puede funcionar con velocidades de reloj entre 27 y 80 Mhz. Las prestaciones para el TMS320C31-60 (60 Mhz) son:

- Tecnología 0.8 μm CMOS.
- Ejecución de instrucciones en 1 ciclo de 33 ns.
- 30 MIPS.
- 60 MFLOPS.
- 2 bloques de memoria RAM internos de 1K x 32 bits de acceso dual en un solo ciclo.
- Espacio de memoria de 2^{24} palabras de 32 bits.
- 64 x 32 bits cache de instrucciones.
- Instrucciones y datos de 32 bits.
- Direcciones de 24 bits.
- ALU y multiplicador en coma flotante/fija de 40/32 bits.
- Desplazador en tonel de 32 bits.
- 8 registros acumuladores de precisión extendida.
- 2 generadores de direcciones con 8 registros auxiliares y 2 unidades aritméticas para registros auxiliares.
- Controlador DMA interno para operaciones de E/S concurrentes con las operaciones de la CPU.
- Operaciones lógicas, con enteros y coma flotante.
- Instrucciones de 2 y 3 operandos.
- Instrucciones de multiplicación y ALU paralelas en un solo ciclo.
- Bucles con bifurcaciones en un solo ciclo.
- Llamadas a subrutina y retornos condicionales.
- Interbloqueo de instrucciones para soporte de multiproceso.
- 2 buses de datos de 32 bits (direcciones de 24 y 13 bits).
- 1 puerto serie soportando transferencia 8, 16, 24 y 32 bits.
- 2 temporizadores de 32 bits.
- 2 registros con flags de propósito general
- 4 interrupciones externas.
- Empaquetado de 132 pines.

Tabla D.1. Tabla de características del DSP TMS320C31-60.

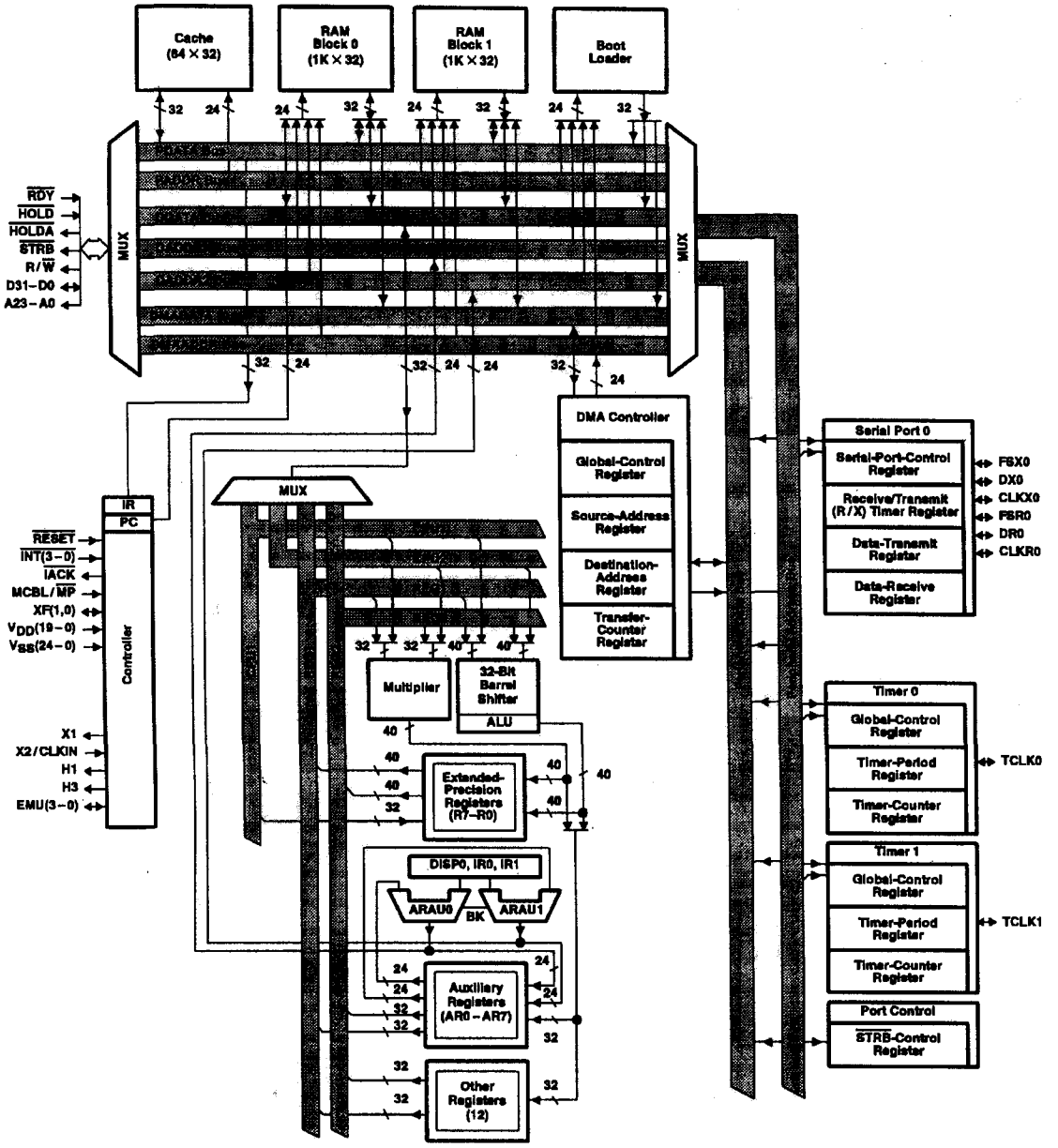

Figura D.1. Diagrama de bloques del TMS320C3x.

D.1.1. Unidad central de procesamiento o CPU

a).- **Multiplicador en coma flotante y enteros.** Realiza en 1 ciclo multiplicaciones de enteros de 24 bits y de reales en coma flotante de 32 bits.

b).- **Unidad aritmético-lógica o ALU (Arithmetic Logic Unit).** Permite realizar operaciones lógicas, con números en coma flotante y fija. Los valores enteros

son de 32 bits y los valores en coma flotante son de 40 bits. Se puede realizar en paralelo una operación de multiplicación y otra donde se emplee la ALU (con determinadas restricciones).

c).- **Unidades aritméticas para registros auxiliares o ARAUs** (*Auxiliary Register Arithmetic Units*). Las 2 unidades presentes (ARAU0 y ARAU1) pueden generar 2 direcciones en un solo ciclo. Las ARAUs operan en paralelo con las unidades de multiplicación y la ALU. Soportan modos de direccionamiento con desplazamiento, registros de indexación (IR0 e IR1), desplazamiento circular y de bit simétrico.

d).- **Buses internos CPU1/CPU2 y REG1/REG2.** Permiten llevar 2 operandos desde la memoria y 2 desde el banco de registros en el mismo ciclo.

e).- **Banco de registros de la CPU.** Posee 28 registros que pueden ser manejados tanto por el multiplicador como por la ALU. Además todos ellos pueden manejarse como registros de propósito general de 32 bits.

f).- Desplazador en tonel de 32 bits.

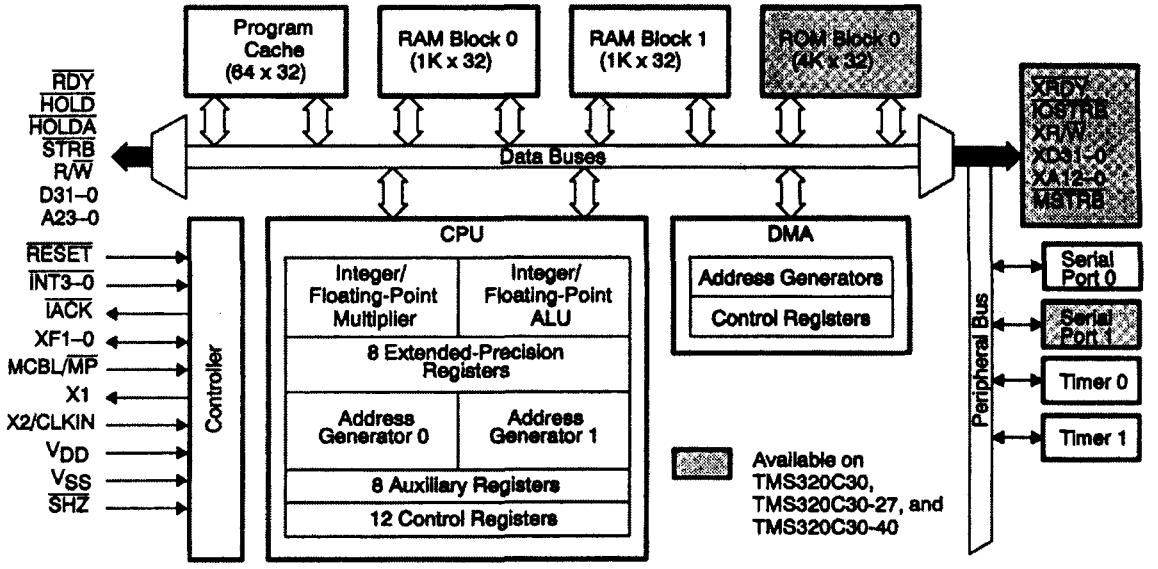

Figura D.2. Detalle de la arquitectura del TMS320C3x.

D.1.2. Mapa de memoria del procesador

El espacio total de direcciones del procesador es de 2^{24} palabras de 32 bits. Programas, datos y espacio de E/S deben estar contenidos en este rango.

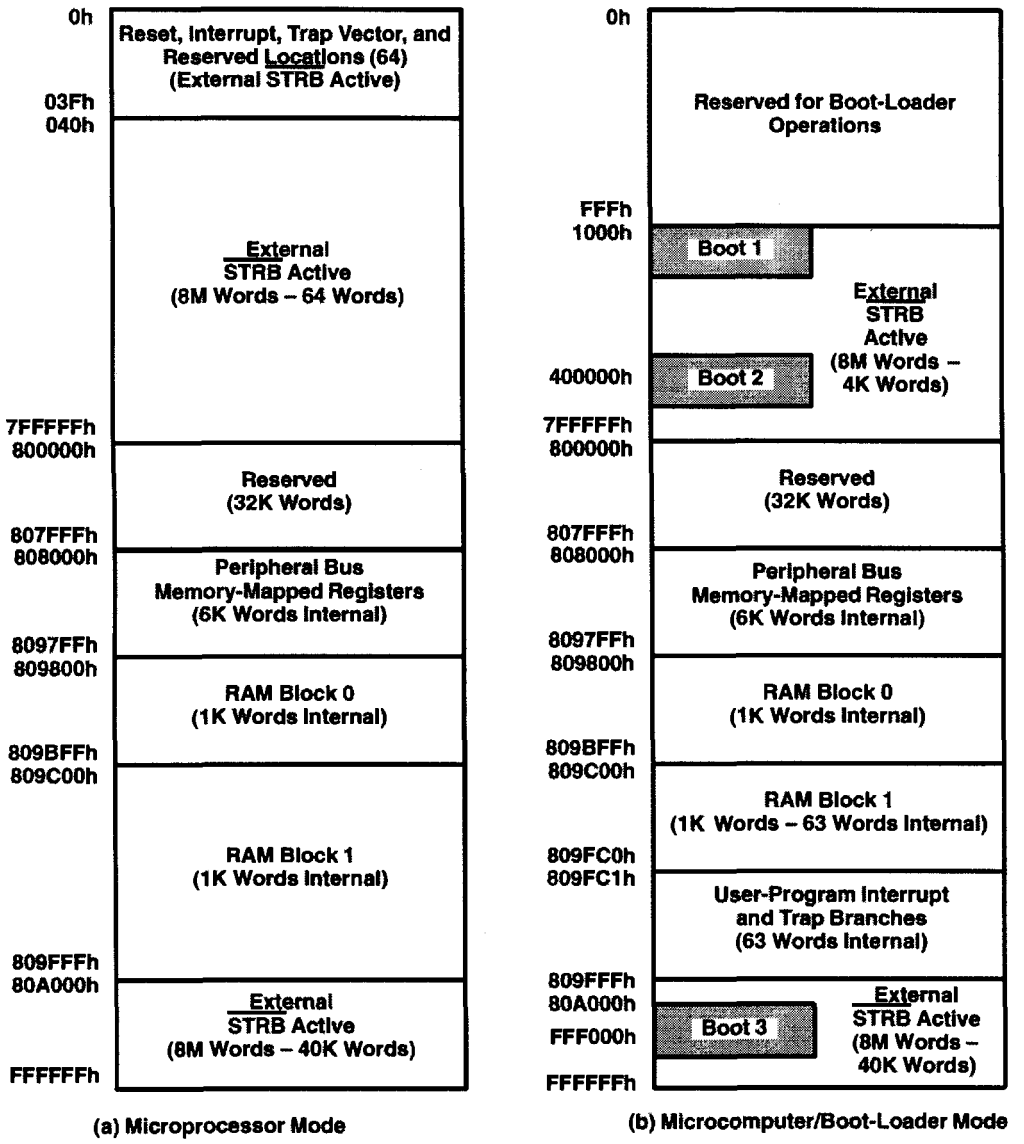

Figura D.3. Mapas de memoria del TMS320C31.

Los 2 bloques de memoria RAM interna se encuentran en las direcciones 809800h-809BFF y 809C00h-809FFFh respectivamente. Cada bloque de RAM es capaz de soportar 2 accesos a la CPU en el mismo ciclo.

El resto del mapa de memoria depende de si el procesador está funcionando en

modo microprocesador o modo microcomputador.

En modo microprocesador no existe la posibilidad de disponer de un cargador (*boot loader*). Las posiciones 0h-0BFh contendrán los vectores de interrupción, *trap* y posiciones reservadas.

En modo microcomputador las posiciones en el rango 0h-0FFFh se reservan para las operaciones con el cargador.

D.1.3. Estructura del pipeline

Presenta cinco unidades que son:

1. **Fetch.** Carga desde la memoria la siguiente instrucción y actualiza el registro contador de programa o PC (*Program Counter*).
2. **Decode.** Decodifica la instrucción y realiza la generación de dirección. Esta unidad controla también cualquier modificación en los registros auxiliares y el puntero de pila.
3. **Read.** Lee los operandos desde la memoria si fuera necesario.
4. **Execute.** Lee los operandos desde el banco de registros, si se requiere, realiza cualquier operación necesaria y escribe los resultados en el banco de registros. Si se necesita, la unidad escribe los resultados de las operaciones previas en memoria.
5. **DMA Channel.** Lee y escribe en memoria.

Las instrucciones básicas tienen 4 niveles que son:

1. Fetch.
2. Decode.
3. Read,
4. Execute.

El orden de prioridad de las distintas unidades es:

Execute (Mayor prioridad).
Read.
Decode.
Fetch.
DMA (Menor prioridad).

Cuando el procesamiento de una instrucción hace que se encuentre preparada para pasar a la siguiente etapa, pero ésta última no pueda admitir la nueva entrada, se produce un conflicto en el *pipeline*. En este caso, la unidad con menor prioridad esperará hasta que la unidad con mayor prioridad complete su actual ejecución.

Debido a la mínima prioridad del controlador de DMA, se puede minimizar o incluso eliminar los conflictos con la CPU, a través de la estructuración de los datos, ya que el controlador posee sus propios buses de direcciones y datos.

D.1.4. Registros de la CPU

Registro	Función
R0- R7	Registros acumuladores extendidos de 32 bits para enteros y 40 bits para coma flotante.
AR0- AR7	Registros de dirección.
DP	Registro de pagina.
IR0- IR1	Registros de indexación.
BK	Registro tamaño de bloque.
SP	Puntero de pila del sistema.
ST	Registro de estado.
IE	Registro de mascaras de interrupción.
IF	Flags de interrupción.
IOF	Flags de entrada/salida.
RS, RE, RC	Registro de inicio, fin y contador para bucles.

Tabla D.2. Registros de la CPU.

D.1.5. Juego de instrucciones del procesador

Instrucción	Descripción	Operación
ABSF	Valor absoluto de un número en coma flotante	|src| → Rn
ABSI	Valor absoluto de un entero	|src| → Dreg
ADDC	Suma de enteros con acarreo	src + Dreg + C → Dreg
ADDC3	Suma de 2 enteros con acarreo (3 operandos)	src1 + src2 + C → Dreg
ADDF	Suma de 2 valores en coma flotante	src + Rn → Rn
ADDF3	Suma de 2 valores en coma flotante (3 operandos)	src1 + src2 → Dreg
ADDI	Suma de 2 enteros	src + Dreg → Dreg
ADDI3	Suma de 2 enteros (3 operandos)	src1 + src2 → Dreg
AND	Operación AND lógica a nivel de bit	Dreg AND src → Dreg
AND3	Operación AND lógica a nivel de bit (3 operandos)	src1 AND $\overline{src2}$ → Dreg
ANDN	Operación AND lógica a nivel de bit con 2° operando negado	Dreg + src → Dreg
ANDN3	Operación AND lógica a nivel de bit con 2° operando negado (3 operandos)	src1 AND $\overline{src2}$ → Dreg
ASH	Desplazamiento aritmético	Si desplazamiento ≥ 0: (Desplazar Dreg a la izquierda desplazamiento → Dreg Caso contrario: (Desplazar Dreg a la derecha |desplazamiento| → Dreg;
ASH3	Desplazamiento aritmético (3 operandos)	Si desplazamiento ≥ 0: (Desplazar src a la izquierda desplazamiento → Dreg Caso contrario: (Desplazar src a la derecha |desplazamiento| → Dreg;
Bcond	Bifurcación condicional (estándar)	Si cond = Verdadera: Si Csrc es registro, Csrc → PC; Si Csrc es valor, Csrc + PC → PC; Caso contrario, PC +1 → PC;
BcondD	Bifurcación condicional (retardada)	Si cond = Verdadera: Si Csrc es registro, Csrc → PC; Si Csrc es valor, Csrc + PC+ 3 → PC; Caso contrario, PC +1 → PC;
BR	Bifurcación incondicional (estándar)	Valor → PC
BRD	Bifurcación incondicional (retardada)	Valor → PC
CALL	Llamada a subrutina	PC +1 → TOS Valor → PC
CALLcond	Llamada a subrutina con cond	Si cond = Verdadera: PC +1 → TOS Si Csrc es registro, Csrc → PC; Si Csrc es valor, Csrc + PC → PC; Caso contrario, PC +1 → PC;
CMPF	Comparación de 2 valores en coma flotante	Actualizar "flags" para Rn – src
CMPF3	Comparación de 2 valores en coma flotante (3 operandos)	Actualizar "flags" para src1 – src2
DBcond	Decremento y bifurcación condicional (estándar)	ARn – 1 → ARn Si cond = Verdadera y ARn ≥ 0: Si Csrc es registro, Csrc → PC; Si Csrc es valor, Csrc + PC + 1 → PC; Caso contrario, PC +1 → PC;
DbcondD	Decremento y bifurcación condicional (retardado)	ARn – 1 → ARn Si cond = Verdadera y ARn ≥ 0: Si Csrc es registro, Csrc → PC; Si Csrc es valor, Csrc + PC + 3 → PC;

		Caso contrario, PC +1 → PC;
FIX	Conversión de un valor en coma flotante en un entero	Fix(src) → Dreg
FLOAT	Conversión de un valor entero en un número en coma flotante	Fix(src) → Dreg
IACK	Reconocimiento de una interrupción	Lectura de scr IACK bajo y luego alto
IDLE	Parada hasta la siguiente interrupción	PC + 1→ PC
LDE	Cargar el exponente en un valor en coma flotante	src(exponente) → Rn(exponente)
LDF	Cargar un valor en un número en coma flotante	src → Rn
LDFcond	Cargar un valor en un número en coma flotante de manera condicional	Si cond = Verdadera: src → Rn
LDFI	Cargar un valor en un número en coma flotante con interbloqueo de señales	src → Rn con señales interbloqueadas
LDI	Cargar un valor entero	src → Dreg
LDIcond	Cargar un valor entero de manera condicional	Si cond = Verdadera: src → Dreg
LDII	Cargar un valor entero con interbloqueo	src → Dreg con señales interbloqueadas
LDM	Cargar la mantisa en un valor en coma flotante	src(mantisa) → Rn(mantisa)
LSH	Desplazamiento lógico	Si desplazamiento ≥ 0: (Desplazar Dreg a la izquierda desplazamiento → Dreg Caso contrario: (Desplazar Dreg a la derecha |desplazamiento| → Dreg;
LSH3	Desplazamiento lógico (3 operandos)	Si desplazamiento ≥ 0: (Desplazar src a la izquierda desplazamiento → Dreg Caso contrario: (Desplazar src a la derecha |desplazamiento| → Dreg;
MPYF	Multiplicación de 2 valores en coma flotante	src x Rn → Rn
MPYF3	Multiplicación de 2 valores en coma flotante (3 operandos)	src1 x src2 → Rn
MPYI	Multiplicación de 2 valores enteros	src x Dreg → Dreg
MPYI3	Multiplicación de 2 valores enteros (3 operandos)	src1 x src2 → Dreg
NEGB	Negación un valor entero empleando el bit de acarreo	0 – src -C → Dreg
NEGF	Negación de un valor en coma flotante	0 – src → Rn
NEGI	Negación de un valor entero	0 – src → Dreg
NOP	Operación nula	Modificar ARn si se especifica
NORM	Normalizar un valor en coma flotante	Normalizar(src) → Rn
NOT	Operación NOT lógica	$\overline{\text{src}}$ → Dreg
OR	Operación OR lógica	Dreg OR src → Dreg
OR3	Operación OR lógica (3 operandos)	src1 OR src2 → Dreg
POP	Extrae un entero de la pila	*SP-- → Dreg
POPF	Extrae un valor en coma flotante de la pila	*SP-- → Rn
PUSH	Introduce un entero de la pila	Sreg → *++SP
PUSHF	Introduce un valor en coma flotante de la pila	Rn → *++SP
RETIcond	Retorno de interrupción condicional	Si cond = Verdadera o no existe: *SP-- → PC 1 → ST(GIE) Caso contrario, continuar;
RETScond	Retorno de subrutina condicional	Si cond = Verdadera o no existe: *SP-- → PC Caso contrario, continuar;
RND	Redondeo de un valor en coma flotante	Redondear(src) → Rn
ROL	Rotación hacia la izquierda	Rotar Dreg hacia la izquierda 1 bit →

		Dreg
ROLC	Rotación hacia la izquierda a través del bit de acarreo	Rotar Dreg hacia la izquierda 1 bit a través de C → Dreg
ROR	Rotación hacia la derecha	Rotar Dreg hacia la derecha 1 bit → Dreg
RORC	Rotación hacia la derecha a través del bit de acarreo	Rotar Dreg hacia la derecha 1 bit a través de C → Dreg
RPTB	Repetición de un bloque de instrucciones	Rotar Dreg hacia la izquierda 1 bit → Dreg
RPTS	Repetición de una instrucción simple	src → RE 1 → ST(RM) PC siguiente → RS PC siguiente → RE
SIGI	Interbloqueo de señales	Operación de interbloqueo de señales Esperar al reconocimiento del interbloqueo Eliminar interbloqueo
STF	Almacenar un valor en coma flotante	Rn → Daddr
STFI	Almacenar un valor en coma flotante con interbloqueo	Rn → Daddr Fin de interbloqueo
STI	Almacenar un valor entero	Sreg → Daddr
STII	Almacenar un valor entero con interbloqueo	Sreg → Daddr Fin de interbloqueo
SUBB	Resta 2 operandos empleando el bit de acarreo	Dreg –src - C → Dreg
SUBB3	Resta 2 operandos empleando el bit de acarreo (3 operandos)	src1 – src2 - C → Dreg
SUBC	Resta 2 enteros condicionalmente	Si Dreg – src ≥ 0: [(Dreg – src) << 1] OR 1 → Dreg Caso contrario, Dreg << 1 → Dreg;
SYBF	Resta 2 valores en coma flotante	Rn – src → Rn
SUBF3	Resta 2 valores en coma flotante (3 operandos)	src1 – src2 → Rn
SUBI	Resta 2 valores enteros	Dreg– src → Dreg
SUBI3	Resta 2 valores enteros (3 operandos)	src1 – src2 → Dreg
SUBRB	Resta 2 valores enteros intercambiando el papel de los operandos y empleando además el bit de acarreo	src – Dreg – C → Dreg
SUBRF	Resta 2 valores en coma flotante intercambiando el papel de los operandos	src – Rn → Rn
SUBRI	Resta 2 valores enteros intercambiando el papel de los operandos	src – Dreg → Dreg
SWI	Interrupción software	Desarrolla la secuencia de interrupción del emulador
TRAPcond	Ejecuta un "trap" condicional	Si cond = Verdadera o no existe: Siguiente PC → *++SP Vector N de "trap" → PC 0 → ST(GIE) Caso contrario, continuar;
TSTB	Chequeo de campos de bits	Dreg AND src
TSTB3	Chequeo de campos de bits (3 operandos)	src1 AND src2
XOR	Operación XOR lógica	Dreg XOR src → Dreg
XOR3	Operación XOR lógica (3 operandos)	src1 XOR src2 → Dreg

Leyenda:

ARn	Registros AR7- AR0
C	Bit de acarreo
cond	Código de condición
Csrc	Modos de direccionamiento para bifurcaciones condicionales
desplazamiento	Valor de desplazamiento (modos de direccionamiento generales)
Dreg	Cualquier registro
PC	Contador de programa
Rn	Registros R7- R0

SP	Puntero de pila
src	Modos de direccionamiento generales
src1	Modos de direccionamiento con 3 operandos
src2	Modos de direccionamiento con 3 operandos
Sreg	Cualquier registro
TOS	Tope de pila

Tabla D.3. Conjunto de instrucciones del TMS320C3x.

D.1.6. Juego de instrucciones paralelas

Instrucción	Descripción	Operación
	Instrucciones paralelas aritméticas y de almacenamiento	
ABSF || STF	Valor absoluto de un número en coma flotante	|src2| → dst1 || src3 → dst2
ABSI || STI	Valor absoluto de un entero	|src2| → dst1 || src3 → dst2
ADDF3 || STF	Suma de 2 valores en coma flotante (3 operandos)	src1 + src2 → dst1 || src3 → dst2
ADDI3 || STI	Suma de 2 enteros	src1 + src2 → dst1 || src3 → dst2
AND3 || STI	Operación AND lógica a nivel de bit (3 operandos)	src1 AND src2 → dst1 || src3 → dst2
ASH3 || STI	Desplazamiento aritmético (3 operandos)	Si desplazamiento ≥ 0: src2 << desplazamiento → dst1 || src3 → dst2 Caso contrario, src2 >> |desplazamiento| → dst1 || src3 → dst2
FIX || STI	Conversión de un valor en coma flotante en un entero	Fix(src2) → dst1 || src3 → dst2
FLOAT || STF	Conversión de un valor entero en un número en coma flotante	src2 → dst1 || src3 → dst2
LDF || STF	Cargar un valor en un número en coma flotante con interbloqueo de señales	src2 → dst1 || src3 → dst2
LDI || STI	Cargar un valor entero	src2 → dst1 || src3 → dst2
LSH3 || STI	Desplazamiento lógico (3 operandos)	Si desplazamiento ≥ 0: src2 << desplazamiento → dst1 || src3 → dst2 Caso contrario, src2 >> |desplazamiento| → dst1 || src3 → dst2
MPYF3 || STF	Multiplicación de 2 valores en coma flotante (3 operandos)	src1 x src2 → dst1 || src3 → dst2
MPYI3 || STI	Multiplicación de 2 valores enteros	src1 x src2 → dst1 || src3 → dst2
NEGF || STF	Negación de un valor en coma flotante	0 - src2 → dst1 || src3 → dst2

NEGI || STI	Negación de un valor entero	0 - src2 → dst1 || src3 → dst2
NOT || STI	Operación NOT lógica	$\overline{src1}$ → dst1 || src3 → dst2
OR3 || STI	Operación OR lógica (3 operandos)	src1 OR src2 → dst1 || src3 → dst2
STF || STF	Almacenar un valor en coma flotante con interbloqueo	src1 → dst1 || src3 → dst2
STI || STI	Almacenar un valor entero	src1 → dst1 || src3 → dst2
SUBF3 || STF	Resta 2 valores en coma flotante (3 operandos)	src1 – src2 → dst1 || src3 → dst2
SUBI3 || STI	Resta 2 valores enteros (3 operandos)	src1 – src2 → dst1 || src3 → dst2
XOR3 || STI	Operación XOR lógica (3 operandos)	src1 XOR src2 → dst1 || src3 → dst2
	Instrucciones paralelas de carga	
LDF || LDF	Cargar un valor en un número en coma flotante	src2 → dst1 || src4 → dst2
LDI || LDI	Cargar un entero	src2 → dst1 || src4 → dst2
	Instrucciones paralelas de multiplicación/suma	
MPYF3 || ADDF3	Multiplicación y suma de valores en coma flotante	op1 x op2 → op3 || op4 + op5 → op6
MPYF3 || SUBF3	Multiplicación y resta de valores en coma flotante	op1 x op2 → op3 || op4 - op5 → op6
MPYI3 || ADDI3	Multiplicación y suma de valores enteros	op1 x op2 → op3 || op4 + op5 → op6
MPYI3 || SUBI3	Multiplicación y resta de valores enteros	op1 x op2 → op3 || op4 - op5 → op6

Leyenda:		
	desplazamiento	Valor de desplazamiento (modos de direccionamiento generales)
	src1	Registros R7- R0
	src2	Direccionamiento indirecto con desplazamientos 0, 1, IR0 o IR1
	src3	Registros R7- R0
	dst1	Registros R7- R0
	dst2	Direccionamiento indirecto con desplazamientos 0, 1, IR0 o IR1
	op1, op2, op4, op5	2 operandos cualesquiera deben ser registros R7- R0 y los otros 2 direccionamiento indirecto con desplazamientos 0, 1, IR0 o IR1
	op3	Registros R0 o R1
	op6	Registros R2 o R3

Tabla D.4. Conjunto de instrucciones paralelas del TMS320C3x.

D.1.7. Modos de direccionamiento

Modos de direccionamiento		Ejemplo
Modos de direccionamiento generales		
Inmediato	Corto. Emplea 16 bits.	SUBI 1,R0
	Largo. Emplea 24 bits.	BR 805Ch
Registro		ABSF R1
Directo (Emplea el registro DP)		ADDI @0BCDEh, R7
Indirecto	Sin desplazamiento.	STF R0, *AR0
	Con desplazamiento de 8 bits mas preincremento, predecremento, postincremento o postdecremento.	STF R0, *+AR0(7h)
	Con desplazamiento indicado por IR0 o IR1.	STF R0, *+AR0(IR1)
	Con desplazamiento de IR0 o IR1 mas preincremento, predecremento, postincremento o postdecremento.	STF R0, *--AR0(IR1)
	Direccionamiento con desplazamiento circular indicado, por IR0 o IR1. El tamaño del bloque estará contenido en el registro BK.	STF R0, *AR0--(IR1)%
	Direccionamiento con desplazamiento de bit simétrico, indicado por IR0 o IR1. El tamaño del bloque estará contenido en el registro BK.	STF R0, *AR0++(IR1)B
Modos de direccionamiento en instrucciones de 3 operandos		
	Incluye los anteriores cuando sean aplicables pero con algunas limitaciones: * Los desplazamientos podrán ser solo de 0 o 1 o indicados por IR0 o IR1. * Solamente 2 operandos podrán contener direccionamientos indirectos.	ADDC3 *AR5++(IR0), R5, R2
Modos de direccionamiento para instrucciones paralelas		
	Añaden nuevas restricciones al caso anterior aunque se empleen instrucciones de 2 operandos.	MPYF3 R5, *++AR7(IR1), R0 || SUBF3 R7, *AR3—(1), R2

Tabla D.5. Modos de direccionamiento del TMS320C3x.

D.2. HERRAMIENTAS DE DESARROLLO

Son 4 las herramientas de desarrollo del entorno del TMS320C3x/4x:

a).- Compilador ANSI C con optimizador de código.

b).- Macroensamblador.

c).- Simulador software.

d).- Emulador hardware.

La estrategia de desarrollo comienza con la implementación en C de todo el código. Aquellas funciones que requieran un nivel de optimización superior o que no puedan optimizarse desde código C deberán realizarse en ensamblador.

D.2.1. Limitaciones del compilador C

1. No permite especificar las instrucciones que deben ser ejecutadas en paralelo. Ej: copiar vectores de datos (Tabla D.6, Tabla D.7).
2. No permite especificar todos los modos de direccionamiento del TMS320C31. Ej: acceso a registro base con desplazamiento (Tabla D.8).
3. No existen operadores para determinadas funciones. Ej: extracción del exponente de un número en coma flotante.
4. No permite construir como funciones "inline" aquellas escritas en ensamblador.

Ejemplo 1(a). Copiar vectores de datos

Programa C

```
int buffer1[1000];
int buffer2[1000];
int *p_buffer1;
int *p_buffer2;
int i;

/* ... */


p_buffer1= buffer1;
p_buffer2= buffer2;
aux= *p_buffer2++;
for(i=998;i>=0;i--)
      *p_buffer1++= *p_buffer2++;
```

Programa Ensamblador

```
LDI       @_buffer1, AR1
LDI       @_buffer2, AR2
LDI       *AR2++, R0
LDI       998, RC
RPTB      finBucle
finBucle: STI   *AR2++, *AR1++
```

¡ Error !

Tabla D.6. Ejemplo de optimización no realizable en ensamblador por limitación en los operandos del modo de direccionamiento indirecto.

Ejemplo 1(b). Copiar vectores de datos

Programa C

```
int buffer1[1000];
int buffer2[1000];
int *p_buffer1;
int *p_buffer2;
int aux;
int i;

/* ... */


P_buffer1= buffer1;
P_buffer2= buffer2;
Aux= *p_buffer2++;
For(i=998;i>=0;i--)
{
   *p_buffer1++= aux;
   aux= *p_buffer2++;
}
*p_buffer1= aux;
```

Programa Ensamblador

```
LDI                 @_buffer1, AR1
LDI                 @_buffer2, AR2
LDI                 *AR2++, R0
LDI                 998, RC

RPTB                FinBucle
STI                 R0, *AR1*AR2++, R0
finBucle:  || LDI   R0, *AR1++
STI                 R0, *AR1++
```

Tabla D.7. Ejemplo de optimización no realizada por el compilador C.

Ejemplo 2. Acceso a registro base con desplazamiento

Programa C

```
int buffer1[32];
int buffer2[32];
int buffer3[32];
int *p_buffer1;
int *p_buffer2;
int *p_buffer3;
int i;

/* ... */

suma= 0;
p_buffer1= buffer1;
p_buffer2= buffer2;
p_buffer3= buffer3;
for(i=31;i>=0;i--)
{
   int d;
   d= i* i;
   suma+= *(p_buffer1+d);
   suma+= *(p_buffer2+d);
   suma+= *(p_buffer3+d);
}
```

Programa Ensamblador

```
            LDI     0, R0
            LDI     @_buffer1, AR1
            LDI     @_buffer2, AR2
            LDI     @_buffer2, AR3
            LDI     31, RC
            RPTB    finBucle
            STI     RC, RC, IR0
            MPYI    *+AR1(IR0), R0
            ADDI    *+AR2(IR0), R0
finBucle:   ADDI    *+AR3(IR0), R0
```

Tabla D.8. Ejemplo de falta de optimización por parte del compilador C en el manejo del direccionamiento indirecto.

D.2.2. El problema de la operación de división

1. El TMS320C31 no posee una instrucción de división y tampoco existe un cálculo de semillas de la división como en el TMS320C4x.
2. La solución software es muy costosa: *36 ciclos* (Tabla D.9).

Cálculo del inverso (Solución software)

```
v = a * 2**e;
x[0] = 1.0 * 2**(-e-1);
for (i = 1; i <= 5; i++)
   x[i] = x[i-1] * (2.0 - v * x[i-1]);
```

Tabla D.9. Algoritmo de Newton-Raphson para el cálculo del inverso de un número.

D.2.3. La transformada rápida de Fourier (FFT)

El TMS320C31 presenta algunas ventajas de diseño que facilitan su ejecución:

- ✓ Modo de direccionamiento indirecto con desplazamiento indicado por registro.
- ✓ Modo de direccionamiento indirecto con desplazamiento indicado por registro base y desplazamiento de bit simétrico indicado en registro de indexación.
- ✓ Uso extensivo de instrucciones paralelas.

La mejora de la versión ensamblador frente al mismo algoritmo pero escrito en C es de un 40%.

D.3. PROYECTO IVORY: ESTRUCTURA DEL SISTEMA CONJUNTO CANCELADOR DE RUIDO/RECONOCEDOR DE VOZ

La Figura D.4 representa el diagrama de módulos del sistema conjunto cancelador de ruido/reconocedor de voz del proyecto IVORY [IVO.html]:

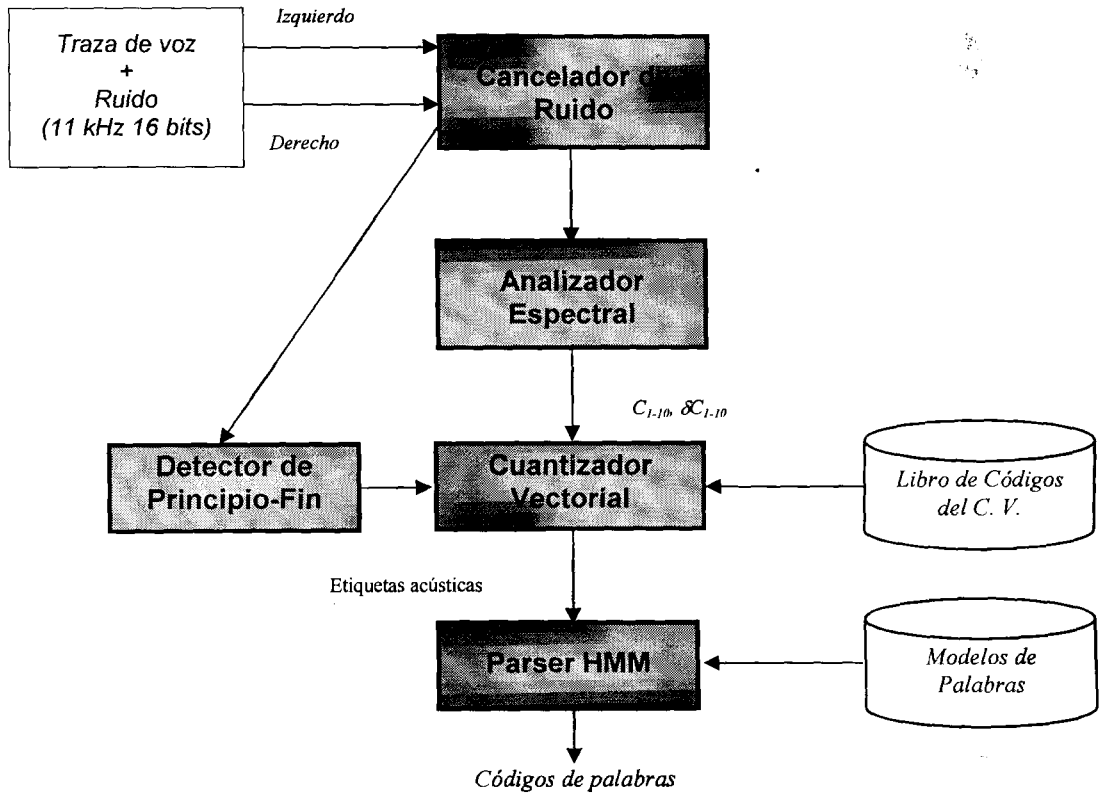

Figura D.4. Esquema de módulos del sistema conjunto cancelador de ruido/reconocedor del proyecto IVORY [IVO.html].

La implementación del sistema (Figura D.5), se realizó empleando un sistema de doble tarjeta como la presentada en la Figura D.6. La parte de cancelación adaptativa se realiza en la primera tarjeta, etiquetada como *NC_HBB*. La cancelación de ruido por substracción espectral y el reconocimiento de voz se ejecuta en la segunda tarjeta, (*SR_HBB*).

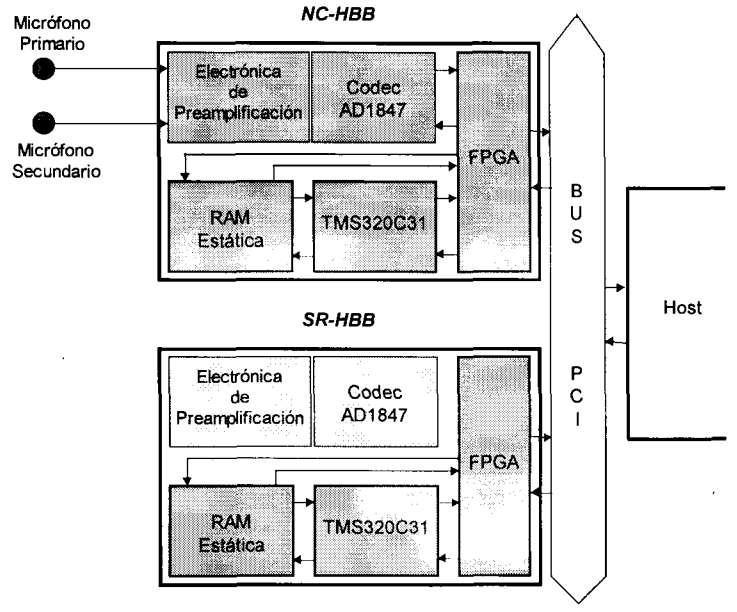

Figura D.5. Estructura del sistema conjunto cancelador de ruido/reconocedor de voz.

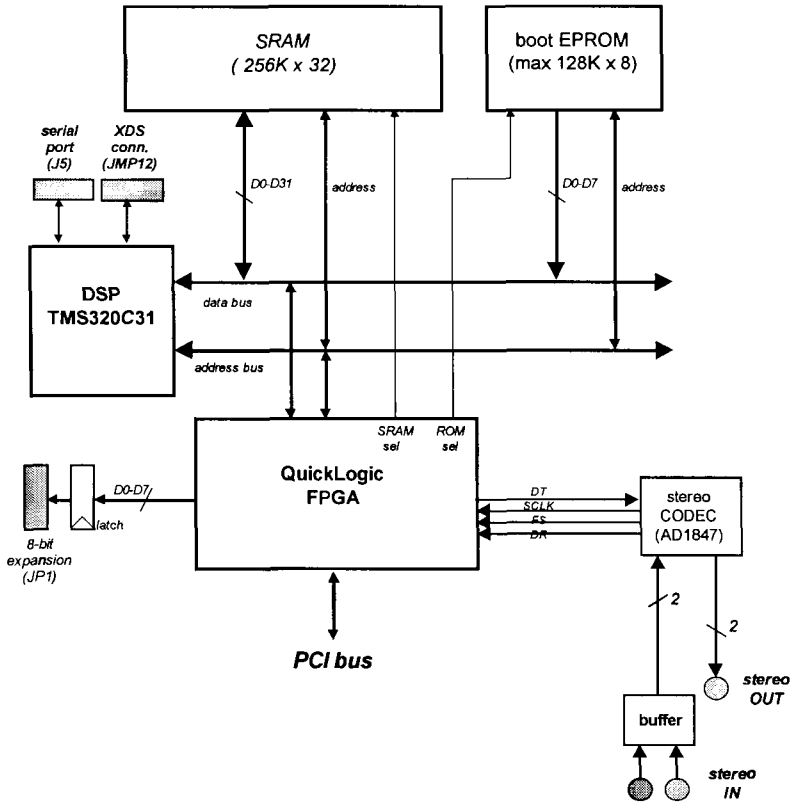

Figura D.6. Diagrama de bloques de la tarjeta diseñada para el proyecto IVORY [IVO.html].

D.3.1. Aspectos que condicionan la implementación del sistema

✓ Dos dimensiones en el estudio: tiempo de ejecución y requisitos de memoria.

✓ Los factores que más condicionan el diseño son los siguientes:

Para el cancelador de ruido:

- Frecuencia de muestreo: 11025Hz.
- Separación entre micrófonos (20cm).
- 7+7 etapas algoritmo LSL.

Para el reconocedor de voz:

- Ciclo básico del reconocedor: 128 muestras.
- Número de índices presentes en los libros de códigos: 256.
- Número de modelos de Markov: 30.

D.3.2. Tarjeta NC-HBB (Implementación del cancelador de ruido)

Pasos:

Por cada muestra de entrada:

- Obtención de muestras de voz.
- Cálculo de la energía de la muestra.
- Cálculo de los cruces por cero.
- Algoritmo Least Square Lattice.
- Almacenamiento de la muestra de voz limpia.

Cada 128 muestras:

- Calcular envolvente de los 2 canales.
- Calcular diferencia de las envolventes de los 2 canales.
- Detección de casos de cruces por cero.
- Autómata de estados: obtención del nuevo peso del filtro LSL.
- Cálculo de la energía de normalización entre canales.
- Cálculo del coeficiente de ajuste entre canales.
- Transferencia de datos.

D.3.3. Tarjeta SR-HBB (Implementación del sustractor espectral y del reconocedor de voz)

Pasos:

Cada 128 muestras:

- Obtención de las muestras de voz (128 muestras).
- Cálculo de la energía con preénfasis de la trama.
- Cálculo de la sustracción espectral.
- Detección de presencia de palabra.
- Desplazar 128 muestras.

Para tramas pertenecientes a una palabra:

- Cálculo de las bandas a partir de la FFT siguiendo escalas de Mel.
- Cálculo de los coeficientes cepstrales.
- Cálculo de los coeficientes cepstrales delta.
- Construcción del vector de parámetros fonéticos.
- Cuantificación vectorial.
- *Parser* de modelos ocultos de Markov.

D.3.4. Tiempos de ejecución

Módulo	Tiempo de Ejecución por Trama	Porción de Tiempo Real
Obtención de muestras	7.424 ciclos	2,13%
Cálculos de energía muestras	5.504 ciclos	1,58%
Cálculo de cruces por cero	4.922 ciclos	1,41%
Algoritmo "Least Square Lattice"	264.704 ciclos	76,00%
Autómata de estados peso	89 ciclos	0,03%
Almacenamiento de muestras	4.608 ciclos	1,32%
Total	**287.251 ciclos**	**82,47%**

Tabla D.10. Tiempos de ejecución para la parte del cancelador de ruido ejecutado en la tarjeta NC-HBB. Los tiempos de ejecución son los correspondientes a una trama de datos de entrada.

Módulo	Tiempo de Ejecución por Trama	Porción de Tiempo Real
Obtención de ventana muestras	6.656 ciclos	1,91%
Cálculo de energía preénfasis	3.912 ciclos	1,12%
Sustracción espectral	100.271 ciclos	28,79%
Detección de principio/fin palabra.	368 ciclos	0,11%
Obtención de bandas FFT	3.954 ciclos	1,14%
Cepstrum	392 ciclos	0,11%
Delta cepstrum	256 ciclos	0,07%
Construcción parámetros fonéticos	84 ciclos	0,02%
Cuantificador vectorial	20.869 ciclos	5,99%
Parser H.M.M.	12.740 ciclos	3,66%
Desplazamiento de muestras	308 ciclos	0,09%
Total	**149.810 ciclos**	**43,01%**

Tabla D.11. Tiempos de ejecución para la parte del cancelador de ruido ejecutado en la tarjeta SR-HBB y el reconocedor de voz operando en la misma plataforma. Los tiempos de ejecución son los correspondientes a una trama de datos de entrada.

D.3.5. Requisitos de memoria

Módulo	Tamaño (palabras de 32 bits)
Módulo principal	46
Obtención de ventana muestras	48
Cálculos energía	27
Cruces por cero	100
Autómata de estados	337
Algoritmo "Least Square Lattice"	277
Almacenamiento de muestras	45
Funciones de biblioteca	40
Manejadores	391
Arranque e inicio	245
Total	**1.556**

Tabla D.12. Requisitos de memoria para almacenar la parte del código del cancelador de ruido que se ejecuta en la tarjeta NC-HBB.

Variables	Tamaño (palabras de 32 bits)
Tabla con inversos	129
Buffer muestras	768
Pila	96
Otras variables	260
Total	**1.253**

Tabla D.13. Requisitos de memoria para almacenar las variables de la parte del sistema cancelador de ruido que se ejecuta en la tarjeta NC-HBB.

Módulo	Tamaño (palabras de 32 bits)
Obtención de ventana muestras	68
Cálculo de energía preénfasis	69
Sustracción espectral	210
Detección de principio/fin palabra.	117
Bandas FFT	32
Cepstrum	61
Delta cepstrum	38
Construcción parámetros fonéticos	24
Cuantificador vectorial	24
Parser H.M.M.	213
Desplazamiento de muestras	72
Funciones de biblioteca	40
Manejadores	391
Arranque e inicio	245
Total	**1.604**

Tabla D.14. Requisitos de memoria para almacenar el código del reconocedor de voz y de la parte del código del cancelador de ruido que se ejecuta en la tarjeta SR-HBB.

Variables	Tamaño (palabras de 32 bits)
Tabla cuantificador vectorial	5376
Tablas FFT	448
Tabla cepstrum	264
Buffer muestras	768
Datos sustracción	768
Datos FFT	1280
Datos cepstrum	208
Datos *parser* H.M.M.	93
Pila	96
Otras variables	135
Total	**9.437**

Tabla D.15. Requisitos de memoria para almacenar las variables del reconocedor de voz y de la parte del cancelador de ruido que se ejecutada en la tarjeta SR-HBB.

Variables	Tamaño (palabras de 32 bits)
Modelos Ocultos de Markov (1 modelo)	1578
Modelos Ocultos de Markov (30 modelos)	47340
Modelos Ocultos de Markov (100 modelos)	157800

Tabla D.16. Requisitos de memoria para almacenar el conjunto de modelos HMM.

D.4. PROYECTO IVORY: ESTRUCTURA DEL SISTEMA RECONOCEDOR DE VOZ

La Figura D.4 representa el diagrama de módulos del sistema reconocedor de voz del proyecto IVORY [IVO.html]:

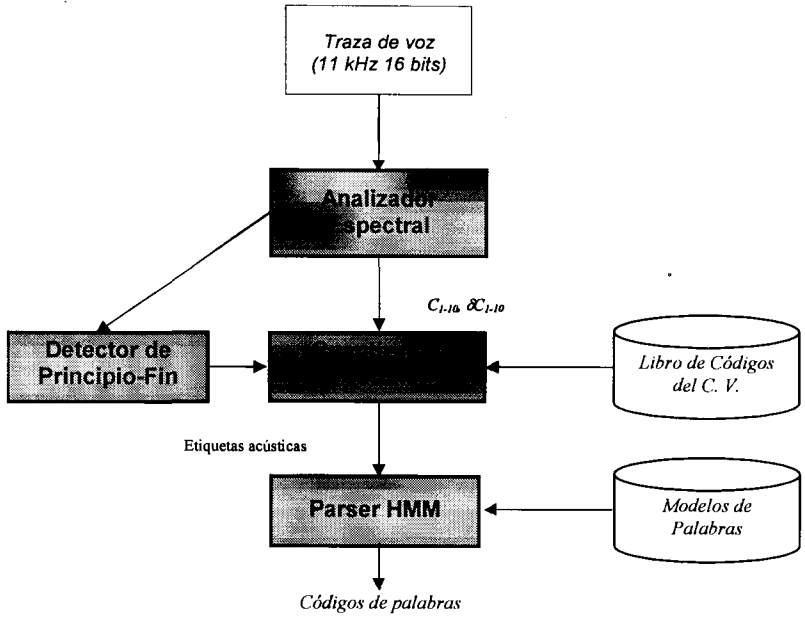

Figura D.7. Esquema de módulos del reconocedor del proyecto IVORY [IVO.html].

La implementación del sistema (Figura D.7), se realiza en este caso empleando una única tarjeta como la presentada en la Figura D.6.

D.4.1. Aspectos que condicionan la implementación del sistema

- ✓ Dos dimensiones en el estudio: tiempo de ejecución y requisitos de memoria.
- ✓ Los factores que más condicionan el diseño son los siguientes:
  - Ciclo básico del reconocedor: 128 muestras.
  - Número de índices presentes en los libros de códigos: 256.
  - Número de modelos de Markov: 30.

D.4.2. Implementación del reconocedor de voz

Pasos del reconocedor basado en FFT-Cepstrum:

Cada trama:

- Obtención de la nueva muestra de voz.
- Cálculo de la energía con preénfasis de la trama.
- Cálculo de la FFT para la trama.
- Cálculo de las bandas a partir de la FFT siguiendo escalas de Mel.
- Detección de presencia de palabra.
- Desplazar 128 muestras.

Para tramas pertenecientes a una palabra:

- Cálculo de los coeficientes cepstrales.
- Cálculo de los coeficientes cepstrales delta.
- Construcción del vector de parámetros fonéticos.
- Cuantificación vectorial.
- *Parser* de modelos ocultos de Markov.

Pasos del reconocedor de voz basado en LPC y cepstrum obtenido a partir de bandas de energía:

Cada muestra:

- Obtención de la nueva muestra de voz.
- Cálculo de la energía con preénfasis de la muestra.
- Ejecutar nueva iteración del algoritmo "Gradient Adaptive Lattice".

Cada trama:

- Cálculo de la transformación de Levinson-Durbin.
- Cálculo de la función de transferencia.
- Cálculo de las bandas a partir del espectro LPC siguiendo escalas de Mel.
- Detección de presencia de palabra.

Para tramas pertenecientes a una palabra:

- Cálculo de los coeficientes cepstrales.
- Cálculo de los coeficientes cepstrales delta.
- Construcción del vector de parámetros fonéticos.
- Cuantificación vectorial.
- *Parser* de modelos ocultos de Markov.

D.4.3. Tiempos de ejecución

Módulo	Tiempo de Ejecución por Trama	Porción de Tiempo Real
Obtención de ventana muestras	6.656 ciclos	1,91%
Cálculo de energía preénfasis	3.912 ciclos	1,12%
FFT	15.621 ciclos	4,48%
Detección de principio/fin palabra.	368 ciclos	0,11%
Obtención de bandas FFT	3.954 ciclos	1,14%
Cepstrum	392 ciclos	0,11%
Delta cepstrum	256 ciclos	0,07%
Construcción parámetros fonéticos	84 ciclos	0,02%
Cuantificador vectorial	20.869 ciclos	5,99%
Parser H.M.M.	12.740 ciclos	3,66%
Desplazamiento de muestras	308 ciclos	0,09%
Total	**65.160 ciclos**	**18,71%**

Tabla D.17. Tiempos de ejecución para el reconocedor de voz usando parámetros del tipo FFT cepstrum. Los tiempos de ejecución son los correspondientes a una trama de datos de entrada.

Módulo	Tiempo de Ejecución por Trama	Porción de Tiempo Real
Obtención de muestras	6.656 ciclos	1,91%
Cálculo de energía preénfasis	3.912 ciclos	1,12%
Algoritmo "Gradient Adaptive Lattice"	64.957 ciclos	18,65%
Detección de principio/fin palabra.	242 ciclos	0,07%
Transformación de Levinson-Durbin	3.060 ciclos	0,88%
Función de transferencia	40.116 ciclos	11,52%
Obtención de bandas energía	3.954 ciclos	1,14%
Cepstrum	392 ciclos	0,11%
Delta cepstrum	256 ciclos	0,07%
Construcción de parámetros fonéticos	84 ciclos	0,02%
Cuantificador vectorial	20.869 ciclos	5,99%
Parser H.M.M.	12.740 ciclos	3,66%
Total	**157.238 ciclos**	**45,14%**

Tabla D.18. Tiempos de ejecución para el reconocedor de voz usando el algoritmo "Gradient Adaptive Lattice" para obtener los coeficientes cepstrales derivados de las bandas de energía. Los tiempos de ejecución son los correspondientes a una trama de datos de entrada.

D.4.4. Requisitos de memoria

Módulo	Tamaño (palabras de 32 bits)
Obtención de ventana muestras	68
Cálculo de energía preénfasis	69
Detección de principio/fin palabra.	8
Bandas FFT	32
Cepstrum	61
Delta cepstrum	38
Construcción parámetros fonéticos	24
Cuantificador vectorial	24
Parser H.M.M.	213
Desplazamiento de muestras	72
Funciones de biblioteca	40
Manejadores	276
Arranque e inicio	245
Total	**1.170**

Tabla D.19. Tarjeta SR-HBB. Requisitos de memoria para almacenar el código del reconocedor de voz basado en parámetros FFT cepstrum.

Variables	Tamaño (palabras de 32 bits)
Tabla cuantificador vectorial	5376
Tablas FFT	448
Tabla cepstrum	264
Buffer muestras	256
Datos FFT	640
Datos cepstrum	208
Datos *parser* H.M.M.	93
Pila	96
Otras variables	121
Total	**7.502**

Tabla D.20. Tarjeta SR-HBB. Requisitos de memoria para almacenar las variables del sistema reconocedor de voz basado en parámetros FFT cepstrum.

Módulo	Tamaño (palabras de 32 bits)
Obtención de ventana muestras	68
Cálculo de energía preénfasis	69
Algoritmo "Gradient Adaptive Lattice"	232
Detección de principio/fin palabra.	8
Transformación de Levinson-Durbin	124
Función de transferencia	363
Obtención de bandas energía	32
Cepstrum	61
Delta cepstrum	38
Construcción parámetros fonéticos	24
Cuantificador vectorial	24
Parser H.M.M.	213
Funciones de biblioteca	40
Manejadores	276
Arranque e inicio	245
Total	**1.817**

Tabla D.21. Tarjeta SR-HBB. Requisitos de memoria para almacenar el código del Reconocedor de Voz basado en el algoritmo "Gradient Adaptive Lattice".

Variables	Tamaño (palabras de 32 bits)
Tabla cuantificador vectorial	5376
Tabla con inversos	129
Tabla cepstrum	264
Tabla de senos y cosenos	256
Buffer muestras	256
Datos cepstrum	208
Datos *parser* H.M.M.	93
Pila	96
Otras variables	244
Total	**6.922**

Tabla D.22. Tarjeta SR-HBB. Requisitos de memoria para almacenar las variables del sistema Reconocedor de Voz basado en el algoritmo "Gradient Adaptive Lattice".